Chapter 4

Elements of Complex Analysis

This chapter presents important concepts from the vast field of complex analysis.

4.1 Functions of one complex variable

Some functions of the complex plane are introduced in this section. These functions arise in a number of applications.

Definition 4.1.1 (Polynomial) *A (scalar) polynomial p in the complex variable s is a function of the form:*

$$p(s) = \sum_{k=0}^{n} p_k s^k \tag{4.1}$$

where the polynomial coefficients $\{p_k\}$ are complex numbers. The degree of p is m if and only if s^m is the largest power with non-zero polynomial coefficient. The zero polynomial has no non-zero coefficient. Its degree is 0. $\star$

A function of the form:

$$\sum_{k=0}^{n} p_k \frac{1}{z^k}$$

is a polynomial in $1/z$. Now, the coefficients can be linear operators in general. For example,

$$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} s + \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} s^2$$

where the coefficients are matrices. Such polynomials are important in multi-variable linear systems theory.

Definition 4.1.2 (Rational function) *A (scalar) rational function r in the variable s is a function of the form:*

$$r(s) = \frac{p(s)}{q(s)} \tag{4.2}$$

where p and $q \neq 0$ are polynomials in s. A rational function is real rational if and only if the coefficients of its numerator and denominator polynomials are real numbers. A rational function is proper if and only if

$$\lim_{s \to \infty} r(s)$$

exists (as a complex number), in which case, the limit is denoted as $r(\infty)$. A proper rational function for which $r(\infty) = 0$ is called strictly proper rational. ★

Example 4.1.1 *Every polynomial is a rational function. The functions*

$$2, \quad \frac{1}{s+1}, \quad \frac{s^2+3}{s+1}, \quad s^3$$

are all real rational functions. The first is proper but not strictly proper, the second is strictly proper and, the last two are not proper. △

Rational functions have many important properties. We shall discuss some of them shortly.

Definition 4.1.3 (Analytic function) *A function is analytic at a point in the complex plane if and only if it is differentiable at that point. A function that is analytic at every point in an open set is said to be analytic in that open set.* ★

Example 4.1.2 *A polynomial p is analytic at $1 \in \mathbb{C}$. In fact, it is analytic at every point in the complex plane. The rational function*

$$f(s) = \frac{s+1}{s-1}$$

is analytic at every point except 1. The function

$$f(s) = \overline{s}$$

is not analytic. △

A non-trivial fact is that a function of a complex variable is differentiable once if and only if it is differentiable infinitely many times. Thus, analytic functions are infinitely differentiable at every point of analyticity. In contrast, there are functions of one real variable that are differentiable once but not twice.

Definition 4.1.4 (Power series) *A (formal) power series in the complex variable s is of the form:*

$$p(s) = \sum_{k=0}^{\infty} p_k s^k \tag{4.3}$$

where $\{p_k\}$ are complex numbers. ★

A formal power series reduces to a polynomial when all but a finite number of the coefficients are zero, i.e., $p_k = 0$ for all $k \geq n$ where n is a positive integer. Polynomials as in (4.1) are finite sums and, hence, can be evaluated to a complex number for any $s \in \mathbb{C}$. A power series on the other hand is an infinite sum and may not be well-defined for some $s \in \mathbb{C}$. For example, the geometric series:

$$1 + s + s^2 + \cdots + s^n + \cdots$$

cannot be evaluated to a complex number at $s = 2$ as the partial sums diverge. The partial sums of the series can be written as:

$$1 + s + s^2 + \cdots + s^n = \frac{1 - s^{n+1}}{1 - s}$$

from which we may conclude that the series converges to

$$\frac{1}{1 - s}$$

for $|s| < 1$. This behavior is true of power series in general as the following theorem of Abel shows.

Theorem 4.1.1 (Abel [1]) *Let p be a formal power series. There exists a number R in $[0, \infty]$, called the* radius of convergence*, with the following properties:*

1. *The series is absolutely convergent for every s with $|s| < R$ and uniformly convergent for every s with $|s| \leq \rho < R$.*
2. *If $|s| > R$ the terms of the series are unbounded and the series diverges.*
3. *In $|s| < R$ the series is an analytic function.* ■

The radius of convergence of the geometric series mentioned above is 1. Within the open unit disc, i.e. for all s with $|s| < 1$, the series sums up to the analytic function $1/(1 - s)$. Statement 3 of Abel's theorem says that a power series is an analytic function within any open disc contained in the region of convergence of the series. The converse is also true. If f is an analytic function defined in an open disc of radius R, then it has a power series expansion in that disc. A special case of this fact is the well-known Taylor series [1].

Theorem 4.1.2 (Taylor series) *If f is analytic in the open disc of radius R centered at 0, then f has the power series representation:*

$$f(s) = f(0) + \frac{f'(0)}{1!}s + \cdots + \frac{f^{(n)}(0)}{n!}s^n + \cdots$$

for all s with $|s| < R$. ■

Our interest in power series comes partly from the following very important definition.

Definition 4.1.5 (Exponential function) *The power series*

$$\sum_{k=0}^{\infty} \frac{1}{k!} s^k$$

converges for each $s \in \mathbb{C}$ (that is, radius of convergence is ∞) and is called the exponential function. It is denoted as e^s. ⋆

e^s is clearly analytic at every point in the complex plane. It is not a rational function. See [1] for detailed study of e^s and power series.

Definition 4.1.6 (Poles and zeros) *Let f be a scalar function of the complex variable s. A point s_0 in the complex plane (including ∞) is a pole of f if and only if*

$$\lim_{s \to s_0} f(s) = \infty$$

A point $s_0 \in \mathbb{C}$ where $f(s_0) = 0$ is called a zero of f. ⋆

Example 4.1.3 *The function $f(s) = s$ has a pole at ∞ and a zero at 0. The rational function*

$$f(s) = \frac{s-1}{(s+1)^2}$$

has a pole at $s = -1$ and a zero at $s = 1$. △

Proposition 4.1.1 *Let r be a proper rational function. Then, r is analytic at ∞ and has a power series expansion about ∞ of the form:*

$$r(s) = \sum_{k=0}^{\infty} \frac{r_k}{s^k}$$

The coefficients $\{r_k\}$ are known as Markov parameters (in systems theory) or Taylor coefficients. ■

4.2 Evaluation of a function at an operator

We now turn to the important concept of evaluating functions at matrices. In the above discussion, the expression $f(s)$ meant that the function f is being evaluated at the complex variable s. Thus, if p is a polynomial in s:

$$p(s) = \sum_{k=0}^{n} p_k s^k,$$

then $p(1+3j)$ is the value of p at the point $1+3j$. To calculate this value, we replace every occurrence of s with $1+3j$:

$$p(1+3j) = \sum_{k=0}^{n} p_k \, (1+3j)^k$$

and evaluate the expression on the right hand side. This evaluation procedure can be extended to square matrices.

Definition 4.2.1 (Polynomial evaluated at a square matrix) *Let p be a scalar polynomial of the complex variable s:*

$$p(s) = \sum_{k=0}^{n} p_k s^k$$

and $A \in \mathbb{C}^{m\times m}$. Then, p evaluated at A, denoted by $p(A)$, is given by:

$$p(A) = \sum_{k=0}^{n} p_k A^k = p_0 I + p_1 A + p_2 A^2 + \cdots + p_n A^n$$

and is a $m \times m$ matrix. ⋆

Our next objective is to apply this evaluation concept to power series and analytic functions. There is, however, a difficulty since power series involves infinite sum and it is not clear when

$$\sum_{k=0}^{\infty} p_k A^k$$

exists as a matrix. Polynomials are finite sums. So, they can be evaluated at any complex number s or any square matrix A. We need the following definition to resolve this difficulty.

Definition 4.2.2 (Spectral radius of a square matrix) *Let A be a square matrix. The spectral radius of A is:*

$$\rho(A) = \max_i |\lambda_i(A)|$$

where $\{\lambda_i(A)\}$ are the eigenvalues of A. That is, the spectral radius is the maximum of the absolute values of the eigenvalues of A. ⋆

Theorem 4.2.1 *Let p be a power series:*

$$p(s) = \sum_{k=0}^{\infty} p_k s^k$$

with radius of convergence $R > 0$. Then,

$$\sum_{k=0}^{\infty} p_k A^k$$

exists as a matrix for any square matrix A whose spectral radius is strictly less than R. ■

The theorem can be proved using Jordan forms. It allows us to make the following definitions.

Definition 4.2.3 (Analytic function evaluated at a square matrix) *Let f be a scalar analytic function with power (Taylor) series expansion:*

$$f(s) = \sum_{k=0}^{\infty} f_k s^k$$

whose radius of convergence is $R > 0$. Then, for any square matrix A whose spectral radius is strictly less than R, f evaluated at A denoted by $f(A)$ is given by:

$$f(A) = \sum_{k=0}^{\infty} f_k A^k$$

and is a square matrix of the same dimension as A. ⋆

Definition 4.2.4 (Exponential function evaluated at a square matrix) *For any square matrix A, the exponential function evaluated at A, denoted by e^A, is given by:*

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k$$

and is a square matrix of the same dimension as A. ⋆

It is very important to note that e^A is not the exponential of the elements of A. The same holds for other analytic functions evaluated at a matrix. The evaluation operation, that is, the act of evaluating a function at a matrix, has many interesting properties. We list some of these below for general analytic functions as well as exponential function.

Theorem 4.2.2 (Properties of evaluation) *Let f be an analytic function with radius of convergence R. Let A and B be square matrices whose spectral radii are strictly less than R. The following statements are true.*

1. *$f(A^*)$ exists and is equal to $f(A)^*$. That is, f evaluated at the complex conjugate transpose of A is equal to the complex conjugate transpose of f evaluated at A (we say that evaluation and conjugate transposing commute).*
2. *Let M be an invertible matrix of the same size as A. Then, $f(MAM^{-1})$ exists and is equal to $Mf(A)M^{-1}$.*
3. *Define:*

   $$C = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$$

   i.e., the block-diagonal matrix whose diagonal blocks are A and B. Then, $f(C)$ exists and

   $$f(C) = \begin{bmatrix} f(A) & 0 \\ 0 & f(B) \end{bmatrix}$$

   that is, f evaluated at a block-diagonal matrix is the block-diagonal matrix obtained by evaluating f at the diagonal blocks. ■

Theorem 4.2.3 (Properties of exponential function) *The following statements are true.*

1. *$e^0 = I$, i.e, the exponential function evaluated at the zero matrix is equal to the identity matrix.*
2. *Let A be a square matrix. Then, e^A is invertible and its inverse is given by e^{-A}.*
3. *Let A and B be square matrices. Then, $e^{A+B} = e^A e^B$ if and only if $AB = BA$. In this case, we also have $e^{A+B} = e^B e^A$.* ■

The definition of $f(A)$ involves a power series in A and, to evaluate $f(A)$, we must calculate the infinite sum. This may not be a good way to compute due to numerical errors. Fortunately, linear algebra provides efficient computational procedures which we discuss next.

Theorem 4.2.4 (Evaluating an analytic function - special cases) *Let f be an analytic function with radius of convergence R and A be a matrix with spectral radius strictly less than R. The following statements are true.*

1. *If A is a diagonal matrix, then $f(A)$ is a diagonal matrix whose diagonal elements are f evaluated at the diagonal elements of A.*
2. *If A is an upper-triangular $n \times n$ Jordan matrix:*

$$A = \begin{bmatrix} \lambda & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 1 & 0 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix} = J_n(\lambda)$$

   then $f(A)$ is the $n \times n$ upper-triangular matrix:

$$f(A) = \begin{bmatrix} f(\lambda) & f'(\lambda) & f''(\lambda)/2! & \cdots & f^{(n-2)}(\lambda)/(n-2)! & f^{(n-1)}(\lambda)/(n-1)! \\ 0 & f(\lambda) & f'(\lambda) & \cdots & f^{(n-3)}(\lambda)/(n-3)! & f^{(n-2)}(\lambda)/(n-2)! \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & f(\lambda) & f'(\lambda) \\ 0 & 0 & 0 & \cdots & 0 & f(\lambda) \end{bmatrix} = f(J_n(\lambda))$$

   where

$$f^k(\lambda)$$

   denotes the kth derivative of f evaluated at λ.

3. *If A is a Jordan matrix of the form:*

$$A = \begin{bmatrix} J_{n_1}(\lambda_1) & 0 \\ 0 & J_{n_2}(\lambda_2) \end{bmatrix}$$

then $f(A)$ is given by:

$$f(A) = \begin{bmatrix} f\left(J_{n_1}\left(\lambda_1\right)\right) & 0 \\ 0 & f\left(J_{n_2}\left(\lambda_2\right)\right) \end{bmatrix}$$

where $f\left(J_{n_i}\left(\lambda_i\right)\right)$ for $i = 1, 2$ are as in Statement 2. ■

Accordingly, if A happens to have the special structure indicated in the theorem, then we can easily evaluate $f(A)$ without actually adding up the terms of the infinite sum that defines $f(A)$. Here is an example.

Example 4.2.1 *Let*

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

which is an upper-triangular Jordan matrix with eigenvalue $\lambda = 0$. Take f to be the geometric series:

$$f(s) = \sum_{k=0}^{\infty} s^k$$

whose radius of convergence is $R = 1$. The spectral radius of A is:

$$\rho(A) = \max_{i=1,2} |\lambda_i(A)| = 0$$

which is strictly less than R. Hence, we can apply statement 2 of Theorem 4.2.4 to get:

$$f(A) = \begin{bmatrix} f(0) & f'(0) \\ 0 & f(0) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

where we used the fact that

$$f'(s) = \sum_{k=1}^{\infty} k s^k,$$

$f(0) = 1$ and $f'(0) = 1$. △

Now, if A does not have the special structure mentioned in Theorem 4.2.4, then a different procedure is needed. For this, we recall the Jordan decomposition from Chapter 3.

Theorem 4.2.5 (Complex Jordan form theorem) *Let A be a $n \times n$ matrix (real or complex). There exists an invertible matrix M such that*

$$A = MJM^{-1}$$

where

$$J = \begin{bmatrix} J_{n_1}(\lambda_1) & 0 & \cdots & 0 \\ 0 & J_{n_2}(\lambda_2) & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & J_{n_m}(\lambda_m) \end{bmatrix}$$

and $J_{n_k}(\lambda_k)$ is a $n_k \times n_k$ upper-triangular Jordan matrix with eigenvalue λ_k. The block-diagonal matrix J is called the Jordan form of A. ■

We conclude this section with the following procedure to evaluate an analytic function at a matrix.

Theorem 4.2.6 (Evaluating an analytic function) *Let f be an analytic function with radius of convergence R and A be a matrix with spectral radius strictly less than R. The following statements are true.*

1. *Suppose that A is diagonalizable. Let M be the matrix whose columns are the linearly independent eigenvectors of A and D be the diagonal matrix whose diagonal elements are the corresponding eigenvalues of A, i.e.*

   $$A = MDM^{-1}$$

   is the eigenvalue-eigenvector decomposition of A. Then,

   $$f(A) = Mf(D)M^{-1}$$

   where $f(D)$ is the function f evaluated at the diagonal matrix D.

2. *Suppose that A is non-diagonalizable. Let J be the upper-triangular complex Jordan form of A and M be the similarity transformation that puts A in its Jordan form, i.e.*

   $$A = MJM^{-1}$$

   Then,

   $$f(A) = Mf(J)M^{-1}$$

   where $f(J)$ is the function f evaluated at the block-Jordan matrix D. ■

A computer program to evaluate an analytic function f at a matrix A does the following sequence of operations:

- Compute the eigenvalues and eigenvectors of A
- Calculate the spectral radius of A
- If the spectral radius is greater than or equal to the radius of convergence of f, then exit saying that $f(A)$ cannot be evaluated
- Otherwise, check if there are n linearly independent eigenvectors where n is the size of A:
  - If so, A is diagonalizable and we apply Statement 1 of Theorem 4.2.6
  - If not, we compute the Jordan form of A and apply Statement 2 of Theorem 4.2.6 along with Statements 2-3 of Theorem 4.2.4

This algorithm is much faster and more accurate even at evaluating polynomial functions.

Example 4.2.2 *Let*

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

which is a symmetric matrix. So, it is diagonalizable and we can write:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1}$$

Thus,

$$e^A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} e^3 & 0 \\ 0 & e^1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1} \quad \textit{and} \quad \sin(A) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \sin(3) & 0 \\ 0 & \sin(1) \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1}$$

Similiarly, we can calculate $\cos(A)$, $\log(A)$, *etc.* △

Chapter 5

Normed Linear Spaces and Banach Spaces

Vector space structure permits addition and scalar multiplication of elements. However, it does not support important concepts such as size of elements and distance between elements. This chapter introduces the necessary machinery.

5.1 Norms and normed linear spaces

We have some intuitive notion of the size of an element. For example, if an element is multiplied by a scalar, then the size of the element should get scaled proportionately. The following definition captures this and other intuitive properties. Throughout this section, vector spaces are defined over $\mathbf{F}$ which can be either $\mathbb{C}$ or $\mathbb{R}$.

Definition 5.1.1 (Norm) *Let V be a vector space. A norm on V, denoted by $\|\cdot\|$, is a function from V into $\mathbb{R}$ with the following properties:*

1. $\|x\| \geq 0$ *for all* $x \in V$.
2. $\|x\| = 0$ *if and only if* $x = 0$
3. $\|\alpha x\| = |\alpha|\|x\|$ *for all* $x \in V$ *and scalars* α
4. $\|x + y\| \leq \|x\| + \|y\|$ *for all* x *and* y *in* V. $\star$

Property 1 says that norm of vector is greater than or equal to zero. So, we could think of a norm as a mapping from V into $[0, \infty)$ instead of $\mathbb{R}$ as in the definition. Property 2 is important. It says that there is one and only one element whose norm is zero, namely the zero element of V. This property is used to prove many results. Property 3 is the effect of scaling mentioned earlier. Property 4 is called *triangle inequality* and states that the norm of a sum is no larger than the sum of the norms.

Example 5.1.1 (p-norms or Holder norms on $\mathbf{F}^n$) *Consider the vector space $\mathbf{F}^n$ with the standard basis $\{e_k\}_{k=1}^n$. Recall that $x \in \mathbf{F}^n$ can be written as:*

$$x = \sum_{k=1}^{n} x_k e_k$$

where the scalars $\{x_k\}$ are called the coordinates of x. For $1 \leq p < \infty$, the p-norm of x is given by:

$$\|x\|_p = \left(\sum_{k=1}^{n} |x_k|^p \right)^{1/p} \tag{5.1}$$

and, for $p = \infty$, the ∞-norm of x is given by

$$\|x\|_\infty = \max_{k=1,2,\cdots,n} |x_k| \tag{5.2}$$

When $p = 2$, the definition (5.1) gives the 2-norm:

$$\|x\|_2 = \left(\sum_{k=1}^{n} |x_k|^2 \right)^{1/2} = \sqrt{\sum_{k=1}^{n} \overline{x}_k x_k} = \left(\begin{bmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \right)^{1/2}$$

and is also known as the Euclidean norm of x. △

It is important to notice that the definition of p-norm of x makes use of the coordinates of x in the standard basis. There is nothing special about the standard basis and any basis may be used to define p-norms although the actual value of $\|x\|_p$ will depend on the basis used.

Example 5.1.2 (Weighted 2-norms on $\mathbf{F}^n$) *As before, consider the vector space $\mathbf{F}^n$ with standard basis $\{e_k\}_{k=1}^n$. Let $w = \{w_k\}_{k=1}^n$ be a set of strictly positive numbers. The w-weighted Euclidean norm of x in $\mathbf{F}^n$ is:*

$$\|x\|_2 = \sqrt{\sum_{k=1}^{n} w_k \overline{x}_k x_k} = \left(\begin{bmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_n \end{bmatrix} \begin{bmatrix} w_1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & w_2 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & w_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \right)^{1/2}$$

This is just a special case of general weighted norms induced by positive definite matrices. △

Example 5.1.3 (p-norms on continuous functions on $[0,1]$) *Consider the real vector space $\mathcal{C}([0,1])$ of continuous functions $f : [0,1] \to \mathbb{R}$. For $1 \leq p < \infty$, the p-norm of f in $\mathcal{C}([0,1])$ is given by:*

$$\|f\|_p = \left(\int_0^1 |f(t)|^p dt \right)^{1/p}$$

and, for $p = \infty$, the ∞-norm of f is given by:

$$\|f\|_\infty = \max_{t\in[0,1]} |f(t)|$$

The ∞-norm is also known as the uniform norm. △

Example 5.1.4 (p-norms on continuous functions with compact support) *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. The support of f is:*

$$\text{support} f = \overline{\{x \in \mathbb{R} : f(x) \neq 0\}}$$

that is, the closure of the set of points where f is non-zero. If support f *is compact (closed and bounded, in this case), then f is said to be compactly supported. For example, the function:*

$$f(t) = \begin{cases} 1 & \text{if } t \in (-1,1) \\ 0 & \text{otherwise} \end{cases}$$

is compactly supported with support $[-1,1]$, whereas the Gaussian function:

$$f(t) = e^{-t^2}$$

is supported on $\mathbb{R}$, and does not have compact support.

Let $\mathcal{C}_c(\mathbb{R})$ be the set of all functions $f : \mathbb{R} \to \mathbb{R}$ with compact support (note that different functions in $\mathcal{C}_c(\mathbb{R})$ may have different supports). It is easily shown that $\mathcal{C}_c(\mathbb{R})$ is a vector space. For $1 \leq p < \infty$, the p-norm of f in $\mathcal{C}_c(\mathbb{R})$ is given by:

$$\|f\|_p = \left(\int_{\mathbb{R}} |f(t)|^p dt\right)^{1/p}$$

and, for $p = \infty$, the ∞-norm of f is given by:

$$\|f\|_\infty = \sup_{t\in\mathbb{R}} |f(t)| = \max_{t\in\mathbb{R}} |f(t)|$$

The ∞-norm is also known as the uniform norm. △

Definition 5.1.2 (Normed linear space) *A normed linear space $(V, \|\cdot\|)$ is a vector space V equiped with a norm $\|\cdot\|$.* ⋆

We have seen some examples of normed linear spaces already. It is very easy to construct more examples using the procedures discussed next. Suppose that $(V, \|\cdot\|_V)$ is a normed linear space and W is a vector space isomorphic to V. Then, $\|\cdot\|_V$ induces a norm on W in a natural way. This is the subject of the following proposition.

Proposition 5.1.1 (Norms induced by isomorphisms) *Let $(V, \|\cdot\|_V)$ be a normed linear space and W be a vector space. Suppose that $F : W \to V$ is an isomorphism between W and V. Then, $\|\cdot\|_W$ defined by:*

$$\|w\|_W = \|F(w)\|_V \quad \text{for all } w \in W$$

is a norm on W. ■

Example 5.1.5 (p-norms or Holder norms on $\mathbb{C}^{m\times n}$) *The above proposition can be used to generate Holder norms on $\mathbb{C}^{m\times n}$. To do this, we identify $\mathbb{C}^{m\times n}$ with $\mathbb{C}^{mn}$ with the isomorphism F:*

$$P = [\, p_1 \quad p_2 \quad \cdots \quad p_n \,] \mapsto \begin{bmatrix} p_1 \\ p_2 \\ . \\ . \\ . \\ p_n \end{bmatrix}$$

where p_i is the ith column of p, i.e, convert a matrix to a vector by stacking columns one below the other (called column-major format in computer programming). Then, apply the Holder norms for vectors defined earlier. Thus,

$$\|P\|_k = \left\| \begin{bmatrix} p_1 \\ p_2 \\ . \\ . \\ . \\ p_n \end{bmatrix} \right\|_k$$

for $k \geq 1$. △

Recall the definition of direct sum of subspaces given in Chapter 2 on Page 19. The next result shows how to make direct sum of normed linear spaces into a normed linear space.

Proposition 5.1.2 (Norms on direct sums of normed linear spaces) *Let $(U, \|\cdot\|_U)$ and $(V, \|\cdot\|_V)$ be normed linear spaces where U and V are subspaces of a vector space W. Define*

$$\|\cdot\|_{UV,p} : U \oplus V \to \mathbb{R}$$

as follows:

$$\left\| \begin{bmatrix} u \\ v \end{bmatrix} \right\|_{UV,p} = \left(\|u\|_U^p + \|v\|_V^p \right)^{1/p}$$

for $1 \leq p < \infty$ and

$$\left\| \begin{bmatrix} u \\ v \end{bmatrix} \right\|_{UV,\infty} = \max \left\{ \|u\|_U, \|v\|_V \right\}$$

for $p = \infty$. Then, $\left(U \oplus V, \|\cdot\|_{UV,p} \right)$ is a normed linear space. ■

Example 5.1.6 *Let $(\mathbb{R}^n, \|\cdot\|_2)$ be the real vector space of column vectors of size n with the Euclidean norm and $(\mathcal{C}([0,1]), \|\cdot\|_2)$ be the real vector space of continuous functions on $[0,1]$ with the 2-norm. The direct sum of $\mathbb{R}^n$ and $\mathcal{C}([0,1])$:*

$$\mathbb{R}^n \oplus \mathcal{C}([0,1]) = \left\{ \begin{bmatrix} x \\ f \end{bmatrix} : x \in \mathbb{R}^n, f \in \mathcal{C}([0,1]) \right\}$$

is composed of elements whose first component is a vector and the second component is a function.

With $p = \infty$, we have

$$\left\| \begin{bmatrix} x \\ f \end{bmatrix} \right\|_{\mathbb{R}^n \oplus \mathcal{C}([0,1]),\infty} = \max\left\{\|x\|_2, \|f\|_2\right\} = \max\left\{ \left(\sum_{k=1}^{n} |x_k|^2\right)^{1/2}, \left(\int_0^1 f(t)^2 dt\right)^{1/2} \right\}$$

as a norm on $\mathbb{R}^n \oplus \mathcal{C}([0,1])$. △

We conclude this section with another way to generate normed linear spaces.

Proposition 5.1.3 (Norms induced by restrictions) *Let $(V, \|\cdot\|)$ be a normed linear space and U be a subspace of V. Define*

$$\|x\|_U = \|x\| \quad \text{for all } x \in U$$

Then, $\|\cdot\|_U$ is called the restriction of $\|\cdot\|$ to U and $(U, \|\cdot\|_U)$ is a normed linear space. ■

5.2 Induced metric, balls and sequences

A norm measures the size of elements whereas a metric measures the distance between elements. These are different concepts in general. A norm always induces a natural metric but a metric may not induce a norm. A metric is needed to define important concepts regarding sequences and functions. These will lead us to Banach spaces in the next section.

Definition 5.2.1 (Metric and metric space) *Let S be a set. A metric d on S is a function $d : S \times S \to \mathbb{R}$ with the following properties:*

1. *$d(x, y) \geq 0$ for all x, y in S*
2. *$d(x, y) = 0$ if and only if $x = y$*
3. *$d(x, y) = d(y, x)$ for all x, y in S*
4. *$d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, z in S*

A metric space (S, d) is a set equiped with a metric. ⋆

Compare this definition with that of norm given on Page 67 and observe the similarities. The main differences are that (i) a norm is a function of a single variable, whereas a metric is a function of two variables, and (ii) the definition of a norm involves vector spaces, whereas metrics can be defined on any set. The following example illustrates the second difference.

Example 5.2.1 *Let $S = \{1, 3\}$. Define $d : S \times S \to \mathbb{R}$ as follows:*

$$d(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{otherwise} \end{cases}$$

d is a metric on S, but S is not a vector space. △

Definition 5.2.2 (Induced metric) *Let $(V, \|\cdot\|)$ be a normed linear space. The metric d induced by the norm $\|\cdot\|$ is defined as:*

$$d(x, y) = \|x - y\|$$

for all x, y in V. ⋆

Accordingly, every normed linear space is a metric space. The converse is not true as the following example shows.

Example 5.2.2 *Consider the vector space $\mathbb{R}$ and its subset $S = [1, 2]$. Define:*

$$d(x, y) = |x - y|$$

for all x, y in S. This is the standard metric on S. But, S is not a vector space since $0 \notin S$. △

Rather than using a new notation d to denote an induced metric we will use the notation for norm.

Definition 5.2.3 (Open and closed balls) *Let $(V, \|\cdot\|)$ be a normed linear space. An open ball of radius r centered at $x \in V$ is the set*

$$\{y \in V : \|x - y\| < r\}$$

i.e, the set of all points that are at a distance strictly less than r from x. When the inequality is non-strict, the set is called a closed ball. ⋆

Example 5.2.3 (Ball in $\mathbb{R}^3$ with Euclidean norm) *The terminology "ball" originates from the fact that the definition results in a sphere in $\mathbb{R}^3$ with the Euclidean norm.* △

Example 5.2.4 (Unit ball in $\mathcal{C}([0, 1])$ with uniform norm) *The uniform norm on the real vector space of continuous functions on $[0, 1]$ is given by:*

$$\|f\|_\infty = \max_{t \in [0,1]} |f(t)|$$

So, the unit ball in $\mathcal{C}([0, 1])$ centered at $\sin(t)$ in this norm is:

$$\begin{aligned} \left\{ f \in \mathcal{C}([0,1]) : \max_{t \in [0,1]} |f(t) - \sin(t)| < 1 \right\} &= \{ f \in \mathcal{C}([0,1]) : |f(t) - \sin(t)| < 1, \ \ \forall t \in [0,1] \} \\ &= \{ f \in \mathcal{C}([0,1]) : -1 < f(t) - \sin(t) < 1, \ \ \forall t \in [0,1] \} \\ &= \left\{ \begin{array}{r} f \in \mathcal{C}([0,1]) : -1 + \sin(t) < f(t) < 1 + \sin(t) \\ \text{for all } t \in [0,1] \end{array} \right\} \end{aligned}$$

where we used the fact that $\sin(t) \geq 0$ for all $t \in [0, 1]$ to derive the last equality. △

We now give two important applications of norms, namely convergence of sequences and continuity of functions. Earlier in this chapter and in the previous chapter, we defined spaces of continuous functions without clarifying what is meant by continuity.

Definition 5.2.4 (Boundedness and norm-convergence of a sequence) *Let $(V, \|\cdot\|)$ be a normed linear space and $\{x_k\}_{k=1}^{\infty}$ be a sequence in V.*

1. *The sequence is bounded if and only if there exists $M < \infty$ such that*

$$\|x_k\| \leq M$$

for all k.

2. *The sequence is said to converge in norm to $x \in V$ if and only if*

$$\|x - x_k\| \to 0$$

as k tends to ∞.

When a sequence converges to a point $x \in V$, we say that x is the limit of the sequence. ★

Example 5.2.5 (A bounded non-converging sequence) *The sequence $\{0, 1, 0, 1, \cdots\}$ is a bounded non-converging sequence on the vector space $\mathbb{R}$ equiped with the standard norm.* △

Definition 5.2.5 (Continuity of functions) *Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be normed linear spaces. Let S be an open subset of V.*

1. *A function $f : S \to W$ is continuous at a point $x_0 \in S$ if and only if, for each $\epsilon > 0$, there exists $\delta > 0$ such that*

$$\text{whenever } \|x - x_0\|_V < \delta \text{ we have } \|f(x) - f(x_0)\|_W < \epsilon$$

2. *A function $f : S \to W$ is continuous in S if and only if it is continuous at every point in S.* ★

Example 5.2.6 (Numbers associated with matrices) *Determinant and trace were not viewed as functions in Chapter 3, but as numbers that can be attached with a square matrix. Here, we ask how these numbers vary as the matrices change.*

Let $\left(\mathbb{C}^{n\times n}, \|\cdot\|_2\right)$ and $(\mathbb{C}, \|\cdot\|_2)$ be respectively normed linear spaces of $n \times n$ matrices and complex numbers with the 2-norm. Define the following functions from $\mathbb{C}^{n\times n}$ into $\mathbb{C}$:

$$f(M) = \text{ determinant of } M$$

$$g(M) = \text{ trace of } M$$

It can be shown that f and g are continuous functions. This means that small changes in the matrix will only cause small changes in determinants and traces (although we dont know how small). △

5.3 Banach spaces

This section introduces the notion of a Banach space which is very important in engineering and science. For example, a lot of the recent work in robust control are in the Banach spaces $\mathcal{H}_\infty(\mathbb{D})$, $\mathcal{L}_1(\mathbb{R})$, etc. We shall define these spaces along with other examples of Banach spaces.

The following definition is very important and lies at the heart of what we think convergence of a sequence ought to be.

Definition 5.3.1 (Cauchy sequence) *Let $(V, \|\cdot\|)$ be a normed linear space and $\{x_k\}_{k=1}^{\infty}$ be a sequence in V. The sequence is called a Cauchy sequence if and only if, for each $\epsilon > 0$, there exists a positive integer N such that*

$$\|x_m - x_n\| < \epsilon$$

for all $m \geq N$ and $n \geq N$. ⋆

Suppose that $\{x_k\}_{k=1}^{\infty}$ is a Cauchy sequence. Then, as we take larger and larger values of k, the distance between elements of the sequence do not become larger. More precisely, given any number $\epsilon > 0$, however small it may be, we can find an index N such that all elements of the form x_{N+k}, $k \geq 0$, are within a distance of ϵ of each other.

Example 5.3.1 *The following sequence of matrices:*

$$\left\{ \begin{bmatrix} 1/k & e^{-k} \\ 0 & 1 \end{bmatrix} \right\}_{k=1}^{\infty}$$

is a Cauchy sequence in $\mathbb{R}^{2\times 2}$ with any of the p-norms. On the other hand, the following sequence of matrices:

$$\left\{ \begin{bmatrix} 1/k & \sin(k) \\ 0 & 1 \end{bmatrix} \right\}_{k=1}^{\infty}$$

is not a cauchy sequence. △

Both sequences in the above example are bounded. This shows that bounded sequences are not necessarily Cauchy sequences. But, the following proposition says that Cauchy sequences are bounded.

Proposition 5.3.1 (Cauchy sequences are bounded) *Let $(V, \|\cdot\|)$ be a normed linear space and $\{x_k\}_{k=1}^{\infty}$ be a Cauchy sequence in V. Then, $\{x_k\}_{k=1}^{\infty}$ is bounded.* ■

This proposition can be easily seen as follows. Given an index (a strictly positive integer) N, the subsequence $\{x_{N+1}, x_{N+2}, \cdots\}$ is called the *tail of the sequence $\{x_1, x_2, \cdots\}$ starting at N*. The *head of the sequence ending at N* is $\{x_1, x_2, \cdots, x_N\}$. Note that head of the sequence has only N elements, whereas

the tail is an infinite sequence. Suppose that $\{x_k\}_{k=1}^{\infty}$ is a Cauchy sequence. Pick any $\epsilon > 0$. Then, by definition, there is a strictly positive integer N such that the elements of the tail of the sequence starting at N are within a distance of ϵ of each other. Put differently, if we take a ball of radius 2ϵ centered at x_N, then it will contain the tail starting at N. Now, the head of the sequence ending at N has only a finite number of elements and so, there is one, say x_l, that has the largest norm. It is easy to verify that the sequence is contained in the ball of radius $\|x_l\| + 2\epsilon$ centered at 0.

To define the next concept, let us consider the set of rational numbers. These are all the real numbers that can be written as the ratio of two integers:

$$\mathcal{Q} = \{x \in \mathbb{R} : x = \frac{n}{m} \text{ where } n, m \text{ are integers and } m \neq 0\}$$

For the moment as we are presently interested in sequences, we overlook the fact that $\mathcal{Q}$ is not a vector space over $\mathbb{R}$. On $\mathcal{Q}$, we can define the metric induced by the absolute value. That is, the distance between two numbers is the absolute value of the difference between them. So, $\mathcal{Q}$ with the absolute value metric is a metric space.

Let us consider the sequence of rational numbers:

$$\{1, 1.4, 1.41, 1.414, 1.4142, 1.41421, \cdots\}$$

This is a Cauchy sequence, and the elements get closer and closer as we go towards the tail-end. But, it does not converge to any number in $\mathcal{Q}$ (it converges in $\mathbb{R}$ to the number $\sqrt{2}$ which is not rational). This example shows that Cauchy sequences in metric spaces may not converge even though our initial feeling about the elements getting closer and closer would suggest otherwise. On the other hand, when viewed as a sequence in $\mathbb{R}$, the above sequence is still a Cauchy sequence and converges to a real number.

Definition 5.3.2 (Complete metric space) *A metric space is complete if every Cauchy sequence converges to a point in the metric space.* ⋆

Thus, in a complete metric space, every Cauchy sequence converges to something in that space. When a metric space is not complete, it may be possible to add on limits of Cauchy sequences and obtain a complete metric space. This process is called *completion*.

Example 5.3.2 (Metric spaces of rational numbers and real numbers) *Let $\mathcal{Q}$ be the metric space of rational numbers with absolute value metric. $\mathcal{Q}$ is not complete. Its completion (with respect to absolute value metric) is $\mathbb{R}$ with absolute value metric. $\mathbb{R}$ with absolute value metric is a complete metric space.* △

Example 5.3.3 (Metric space that is not complete) *Consider the normed linear space of $\mathcal{C}([0,1])$ with 2-norm. We have seen earlier that normed linear spaces are metric spaces. So, $\mathcal{C}([0,1])$ with 2-norm is a metric space. It is not complete as there are Cauchy sequences of continuous functions that do not converge to a continuous function.*

Definition 5.3.3 (Banach Space) *A complete normed linear space is called a Banach space.* ⋆

Banach spaces are very important and appear everywhere. We shall now list some important examples and results. For clarity, the rest of this section is divided into finite dimensional examples, spaces of sequences, Lebesgue spaces and Hardy spaces. It should be noted that many other equally important examples (such as Sobolev spaces) are not described. As a reminder, all vector spaces considered in this chapter are defined over either $\mathbb{R}$ or $\mathbb{C}$.

5.3.1 Finite dimensional spaces

In general, to prove that a certain vector space is Banach, we need to show completeness among other things. This is not required if the vector space is finite dimensional as the following theorem states. Recall that a vector space is said to be finite dimensional if and only if it has a basis containing only a finite number of elements.

Theorem 5.3.1 *A finite dimensional normed linear space is a Banach space.* ■

Thus, all the finite dimensional normed linear spaces given previously are Banach spaces.

5.3.2 Spaces of sequences

Let $\mathcal{Z}_+$ denote the set of positive integers and $S(\mathcal{Z}+)$ denote the set of all (one-sided) sequences of real numbers:

$$\begin{aligned} S(\mathcal{Z}+) &= \{(x_1, x_2, x_3, \cdots) : x_k \in \mathbb{R}\} \\ &= \{[x_k]_{k=1}^{\infty} : x_k \in \mathbb{R}\} \end{aligned}$$

We can render this set a vector space by introducing point-wise addition and scalar multiplication.

Definition 5.3.4 (Space of sequences) *For $1 \leq p < \infty$, define the space of sequences $l_p(\mathcal{Z}+)$ as:*

$$l_p(\mathcal{Z}+) = \left\{ [x_k]_{k=1}^{\infty} \in \mathcal{S} : \sum_{k=1}^{\infty} |x_k|^p < \infty \right\}$$

and for $p = \infty$, define the space of sequences $l_\infty(\mathcal{Z}+)$ as:

$$l_\infty(\mathcal{Z}+) = \left\{ [x_k]_{k=1}^{\infty} \in \mathcal{S} : \sup_{k=1,2,\cdots} |x_k| < \infty \right\}$$

⋆

Definition 5.3.5 (p-norms on spaces of sequences) *For $1 \leq p < \infty$, define the p-norm of $x = [x_k]_{k=1}^{\infty} \in l_p(\mathcal{Z}+)$ as:*

$$\|x\|_p = \left(\sum_{k=1}^{\infty} |x_k|^p \right)^{1/p}$$

and for $p = \infty$, define the ∞-norm of $x \in l_\infty(\mathcal{Z}+)$ as:

$$\|x\|_\infty = \sup_{k=1,2,\cdots} |x_k|$$

These norms are knowns as lp-norms. ⋆

Theorem 5.3.2 (l_p spaces are Banach spaces) *For $1 \leq p \leq \infty$, l_p space with the associated lp-norm is a Banach space.* ■

5.3.3 Lebesgue spaces

Lebesgue spaces are usually defined by first introducing Lebesgue measure. As measure-theoretic considerations will take us too far away from the main theme, we state the definition of Lebesgue spaces and introduce associated norms.

Definition 5.3.6 (Lebesgue spaces of functions on $\mathbb{R}$) *For $1 \leq p < \infty$, the Lebesgue space $\mathcal{L}_p(\mathbb{R})$ is given by:*

$$\mathcal{L}_p(\mathbb{R}) = \left\{ f : \mathbb{R} \to \mathbb{R} : f \text{ is Lebesgue measurable and } \int_{\mathbb{R}} |f(t)|^p \, dt < \infty \right\}$$

and for $p = \infty$, the Lebesgue space $\mathcal{L}_\infty(\mathbb{R})$ is given by:

$$\mathcal{L}_\infty(\mathbb{R}) = \left\{ f : \mathbb{R} \to \mathbb{R} : f \text{ is Lebesgue measurable and } \operatorname{ess\,sup}_{t \in \mathbb{R}} |f(t)| < \infty \right\}$$

where the integral is the Lebesgue integral, and ess sup *stands for essential supremum.* ⋆

Definition 5.3.7 (p-norms on Lebesgue spaces) *For $1 \leq p < \infty$, the p-norm of f in the Lebesgue space $\mathcal{L}_p(\mathbb{R})$ is:*

$$\|f\|_p = \left(\int_{\mathbb{R}} |f(t)|^p \, dt \right)^{1/p}$$

and for $p = \infty$, the ∞-norm of f in $\mathcal{L}_\infty(\mathbb{R})$ is:

$$\|f\|_\infty = \operatorname{ess\,sup}_{t \in \mathbb{R}} |f(t)|$$

These norms are called Lp-norms. ⋆

Theorem 5.3.3 (Lebesgue spaces are Banach) *For $1 \leq p \leq \infty$, $\mathcal{L}_p(\mathbb{R})$ with the associated Lp-norm is a Banach space.*

The Lebesgue spaces $\mathcal{L}_1(\mathbb{R})$, $\mathcal{L}_2(\mathbb{R})$ and $\mathcal{L}_\infty(\mathbb{R})$ are very important in engineering. *These are known respectively as the Lebesgue spaces of integrable functions, square integrable functions and essentially bounded functions*. $\mathcal{L}_2(\mathbb{R})$ is also known as *the space of signals of finite energy*. $\mathcal{L}_\infty(\mathbb{R})$ is also known as *the space of persistently exciting signals*.

We now examine $\mathcal{L}_p(\mathbb{R})$ for $p < \infty$ as completions of certain nice spaces. Consider the real vector space $\mathcal{C}_c(\mathbb{R})$ of compactly supported continuous functions $f : \mathbb{R} \to \mathbb{R}$ defined in (5.1.4). Earlier, we turned $\mathcal{C}_c(\mathbb{R})$ into a normed linear space by defining p-norms (or Holder norms):

$$\|f\|_p = \left(\int_{\mathbb{R}} |f(t)|^p \, dt\right)^{1/p}$$

for $1 \leq p < \infty$ and

$$\|f\|_\infty = \sup_{t \in \mathbb{R}} |f(t)|$$

for $p = \infty$. These are precisely the Lebesgue norms because, for continuous functions, Lebesgue integrals become Riemann integrals and essential supremum becomes supremum.

Theorem 5.3.4 (Completion of $\mathcal{C}_c(\mathbb{R})$ in p-norms) *For $1 \leq p < \infty$, the completion of $\mathcal{C}_c(\mathbb{R})$ in the p-norm is $\mathcal{L}_p(\mathbb{R})$.* ■

$\mathcal{L}_\infty(\mathbb{R})$ is the odd space as it is not the completion of $\mathcal{C}_c(\mathbb{R})$ in the ∞-norm (uniform norm). In fact, the completion is a proper subspace of $\mathcal{L}_\infty(\mathbb{R})$.

5.3.4 Hardy spaces

The open unit disc:

$$\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$$

and the open right half plane:

$$\Re = \{z \in \mathbb{C} : \text{ real part of } z > 0\}$$

are the most common subsets of $\mathbb{C}$ arising in engineering applications. They define the region of analyticity of transfer functions of stable systems in discrete-time and continuous-time respectively (some engineers prefer the complement of $\mathbb{D}$ as the region of analyticity). We shall define Hardy spaces of functions on the disc, but the definition can be adapted to the right half plane and, indeed, to any open subset of $\mathbb{C}$.

Definition 5.3.8 (Hardy spaces of the unit disc) *For $1 \leq p < \infty$, the Hardy space $\mathcal{H}_p(\mathbb{D})$ is defined as:*

$$\mathcal{H}_p(\mathbb{D}) = \left\{ f : \mathbb{C} \to \mathbb{C} \, \textit{analytic in } \mathbb{D} \;\; \textit{and} \;\; \sup_{0<r<1} \int_0^{2\pi} |f(re^{i\theta})|^p d\theta < \infty \right\}$$

and for $p = \infty$, the Hardy space $\mathcal{H}_\infty(\mathbb{D})$ is defined as:

$$\mathcal{H}_\infty(\mathbb{D}) = \left\{ f : \mathbb{C} \to \mathbb{C} \text{ analytic in } \mathbb{D} \text{ and } \sup_{0<r<1} \sup_{z:|z|=r} |f(re^{i\theta})| < \infty \right\}$$

$\star$

Definition 5.3.9 (p-norms on Hardy spaces) *For $1 \leq p < \infty$, the p-norm on Hardy space $\mathcal{H}_p(\mathbb{D})$ is defined as:*

$$\|f\|_p = \left(\frac{1}{2\pi} \sup_{0<r<1} \int_0^{2\pi} |f(re^{i\theta})|^p d\theta \right)^{1/p}$$

and for $p = \infty$, the ∞-norm on $\mathcal{H}_\infty(\mathbb{D})$ is given by:

$$\|f\|_\infty = \text{ess} \sup_{0<\theta<2\pi} |f(e^{i\theta})|$$

These norms are known as Hp-norms. $\star$

Theorem 5.3.5 (Hardy spaces are Banach) *For $1 \leq p \leq \infty$, $\mathcal{H}_p(\mathbb{D})$ with the associated Hp-norm is a Banach space.* ■

The spaces $\mathcal{H}_1(\mathbb{D})$, $\mathcal{H}_2(\mathbb{D})$ and $\mathcal{H}_\infty(\mathbb{D})$ are the most important in discrete-time engineering problems. We can define Hardy spaces of analytic functions of the open right half plane analogously for continuous-time problems. While the definition of Hp-norms appear formidable, great simplifications occur when we look at rational functions. In this case, these norms can be calculated from *Bode magnitude* plot. Recall that the Bode magnitude plot of a function f of the complex variable is the plot of

$$\text{frequency } \omega \text{ vs } |f(j\omega)|$$

If, instead of the magnitude, we plot

$$\text{frequency } \omega \text{ vs } |f(j\omega)|^2$$

we get the power spectral plot of f.

Proposition 5.3.2 (Bode plots and Hp norms) *Let f be a rational function analytic in the open right half plane. Let $\Re$ denote the open right half plane. The following statements are true.*

1. *f is in the Hardy space $\mathcal{H}_1(\Re)$ if and only if the area under the Bode magnitude plot of f is finite.*
2. *f is in the Hardy space $\mathcal{H}_2(\Re)$ if and only if the area under the power spectral plot of f is finite.*
3. *f is in the Hardy space $\mathcal{H}_\infty(\Re)$ if and only if the maximum magnitude in the Bode plot is finite.* ■

Chapter 6

Inner Product Spaces and Hilbert Spaces

6.1 Inner products

The dot product of two vectors is the product of their lengths and the cosine of the angle between the vectors. We have a good mental picture of angle between vectors in $\mathbb{R}^2$. The same cannot be said about the angle between functions in infinite dimensional spaces. Inner products generalize the concept of dot product and through it, we can give precise definition of angle between elements of any vector space.

Definition 6.1.1 (Inner product, inner product space) *Let V be a vector space over $\mathbb{C}$. A mapping $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{C}$ satisfying:*

1. *Positivity: $\langle x, x \rangle \geq 0$ for all $x \in V$.*
2. *Positive definite-ness: $\langle x, x \rangle = 0$ if and only if $x = 0$.*
3. *Linearity in the first argument:*

$$\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$$

   for all x, y, z in V and scalars α, β.

4. *Hermitian-ness: $\langle x, y \rangle = \overline{\langle y, x \rangle}$ for for all x, y in V.*

is called an inner product on V.

V together with $\langle \cdot, \cdot \rangle$ is called an inner product space. ⋆

Inner product is a mapping from $V \times V$ into the underlying field. When the underlying field is $\mathbb{C}$, properties 3 and 4 imply that the inner product is *conjugate linear* in the second argument, that is,

$$\langle x, \alpha y + \beta z \rangle = \overline{\langle \alpha y + \beta z, x \rangle}$$

$$\begin{aligned} &= \overline{\alpha\langle y,\, x\rangle + \beta\langle z,\, x\rangle} \\ &= \overline{\alpha}\overline{\langle y,\, x\rangle} + \overline{\beta}\overline{\langle z,\, x\rangle} \\ &= \overline{\alpha}\langle x,\, y\rangle + \overline{\beta}\langle x,\, z\rangle \end{aligned}$$

so that we get the conjugates of α and β instead of α and β. Now, for a real vector space V, an inner product maps $V \times V$ into $\mathbb{R}$. In this case, the inner product is linear in both arguments and the Hermitian-ness in property 4 becomes symmetry.

Example 6.1.1 (Standard inner product on $\mathbb{R}^n$, $\mathbb{C}^n$) *Consider the vector space $\mathbb{R}^n$. Define the following map on $\mathbb{R}^n \times \mathbb{R}^n$ into $\mathbb{R}$:*

$$\langle x,\, y\rangle_{\mathbb{R}^n} = y^T x$$

for all x, y in $\mathbb{R}^n$. It is easy to verify that this is an inner product on $\mathbb{R}^n$. It is the standard inner product on $\mathbb{R}^n$. Note that the definition yields:

$$\langle x,\, x\rangle_{\mathbb{R}^n} = x^T x = \|x\|_2^2$$

that is, the inner product of x and x is the square of the Euclidean norm of x.

The standard inner product on $\mathbb{C}^n$ is given by:

$$\langle x,\, y\rangle_{\mathbb{C}^n} = y^* x$$

for all x, y in $\mathbb{C}^n$. We again get the square of the Euclidean norm as the inner product of x and x. △

Example 6.1.2 (Standard inner product on $\mathbb{C}^{m\times n}$) *Define the following map on $\mathbb{C}^{m\times n} \times \mathbb{C}^{m\times n}$ into $\mathbb{C}$:*

$$\langle A,\, B\rangle_{\mathbb{C}^{m\times n}} = \mathrm{Tr}\,(AB^*)$$

for all A, B in $\mathbb{C}^{m\times n}$. Again, it is easy to verify that this is an inner product. It is the standard inner product on $\mathbb{C}^{m\times n}$. Note that:

$$\langle A,\, A\rangle_{\mathbb{C}^{m\times n}} = \mathrm{Tr}\,(AA^*) = \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2 = \|A\|_2^2$$

that is, the inner product of A and A is the square of the Holder norm $\|A\|_2$. △

Example 6.1.3 (Standard inner product on $\mathcal{L}_2(\mathbb{R})$) *Recall the definition of the Lebesgue space $\mathcal{L}_2(\mathbb{R})$ from Chapter 5. Define the following map:*

$$\langle f,\, g\rangle_{\mathcal{L}_2(\mathbb{R})} = \int_{\mathbb{R}} f(t)g(t)dt$$

for all f, g in $\mathcal{L}_2(\mathbb{R})$. This is the standard inner product on $\mathcal{L}_2(\mathbb{R})$. As before, we have:

$$\langle f,\, f\rangle_{\mathcal{L}_2(\mathbb{R})} = \int_{\mathbb{R}} f(t)^2 dt = \|f\|_2^2$$

which shows that the inner product of f and f is the square of the $\mathcal{L}_2$ norm of f. △

The above examples allude to an important property. An inner product on a vector space induces a norm. Thus, every inner product space is also a normed linear space with the induced norm. Recall that a norm induces a natural metric and that a normed linear space is a metric space. We summarize these below.

Proposition 6.1.1 (Norm induced by an inner product) *Let V be an inner product space with inner product $\langle \cdot, \cdot \rangle$. Define $\|\cdot\| : V \to [0, \infty)$ as follows:*

$$\|x\| = (\langle x, x \rangle)^{1/2} \tag{6.1}$$

for all x in V. Then, $\|\cdot\|$ is a norm on V. ■

The norm defined in the proposition is called the *norm induced by the inner product*. With this norm, V is a normed linear space. The difficult part in proving the proposition is showing triangle inequality. It is an important fact that the definition of an inner product is sufficient for triangle inequality to hold. Inner products also imply other inequalities some of which are summarized below.

Proposition 6.1.2 (Properties of inner products) *Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space. As in Proposition 6.1.1, let $\|\cdot\|$ be the induced norm. The following statements are true.*

1. *Cauchy-Schwarz inequality holds:*

   $$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$$

   for all x, y in V.

2. *Polarization identity holds:*

   $$\langle x, y \rangle = \frac{1}{4} \langle x + y, x + y \rangle - \frac{1}{4} \langle x - y, x - y \rangle$$

   for all x, y in V.

3. *Triangle inequality holds:*

   $$\|x + y\| \leq \|x\| + \|y\|$$

   for all x, y in V. ■

We conclude this section with a few more examples of inner products.

Example 6.1.4 (Weighted inner products on $\mathbb{R}^n$) *Let $A \in \mathbb{R}^m n$ be a matrix with rank n. Define the following map:*

$$\langle x, y \rangle_A = \langle Ax, Ay \rangle_{\mathbb{R}^m} = y^T \left(A^T A \right) x$$

for all x, y in $\mathbb{R}^n$. The map so defined is an inner product on $\mathbb{R}^n$. It induces a weighted 2-norm on $\mathbb{R}^n$. △

Example 6.1.5 (A norm that is not induced by an inner product) *Consider the vector space* $\mathcal{C}([0,1])$ *of continuous functions on* $\mathbb{R}$ *with the uniform-norm:*

$$\|f\|_\infty = \max_{t\in[0,1]} |f(t)|$$

This is a normed linear space. But, the uniform-norm is not induced by an inner product. △

6.2 Hilbert spaces

As we have seen, every inner product space can be made into a normed linear space using the norm induced by the inner product. This permits us to consider sequences, convergence, Cauchy sequences and completeness on inner product spaces as was done in Chapter 5. The culmination of all those concepts is the following.

Definition 6.2.1 (Hilbert space) *A complete inner product space is called a Hilbert space.* ⋆

Recall that a complete normed linear space is called a Banach space. Since every inner product space is also a normed linear space, *Hilbert spaces are Banach spaces*. Some examples of Hilbert spaces are given next.

Example 6.2.1 (finite dimensional examples) $\mathbb{R}^n$ *with the standard inner product:*

$$\langle x,\, y\rangle = y^T x$$

is a Hilbert space. The norm induced by this inner product is the Euclidean norm. Other finite dimensional examples include $\mathbb{C}^n$, $\mathbb{C}^{m\times n}$ *with their respective standard inner products.* △

Example 6.2.2 (Sequence space $l_2(\mathcal{Z}_+)$**)** *The real one-sided sequence space:*

$$l_2(\mathcal{Z}_+) = \left\{ [x_k]_{k=1}^\infty : \sum_{k=1}^\infty x_k^2 < \infty \right\}$$

with the inner product:

$$\langle [x_k]_{k=1}^\infty,\, [y_k]_{k=1}^\infty \rangle = \sum_{k=1}^\infty x_k y_k$$

is a Hilbert space. △

Example 6.2.3 (Lebesgue space of finite energy signals $\mathcal{L}_2(\mathbb{R})$**)** *The Lebesgue space of square integrable functions on* $\mathbb{R}$ *with the inner product:*

$$\langle f,\, g\rangle = \int_{\mathbb{R}} f(t)g(t)dt$$

is a Hilbert space. △

Example 6.2.4 (Hardy space of square integrable analytic functions $\mathcal{H}_2(\mathbb{D})$) *The Hardy space of functions that are analytic in the unit disc and square integrable on the unit circle with the inner product:*

$$\langle f,\, g\rangle = \frac{1}{2\pi}\int_0^{2\pi} f(e^{i\theta})\overline{g(e^{i\theta})}d\theta$$

is a Hilbert space. △

A vector space isomorphism between a normed linear space V and a vector space W induces a norm on W turning it into a normed linear space. An analogous result for Hilbert spaces is the following.

Proposition 6.2.1 (Inner product induced by vector space isomorphism) *Let V be an inner product space with inner product $\langle\cdot,\,\cdot\rangle_V$. Let W be a vector space and $F : V \to W$ be a vector space isomorphism (that is, a linear one-to-one onto map). Define the following map on $W \times W$:*

$$\langle x,\, y\rangle_W = \left\langle F^{-1}(x),\, F^{-1}(y)\right\rangle_V$$

for all x, y in W. Then, $\langle\cdot,\,\cdot\rangle_W$ is an inner product on W. ■

It is important to note that the value of the W-inner product of x and y is equal to the value of the V-inner product of the pre-images of x and y. This suggests the following definitions.

Definition 6.2.2 (Inner product isomorphism) *Let V and W be inner product spaces with inner products $\langle\cdot,\,\cdot\rangle_V$ and $\langle\cdot,\,\cdot\rangle_W$ respectively. A map $F : V \to W$ is an inner product isomorphism if and only if*

1. *F is a vector space isomorphism, and*
2. *$\langle x,\, y\rangle_W = \left\langle F^{-1}(x),\, F^{-1}(y)\right\rangle_V$ for all x, y in W* ★

Definition 6.2.3 (Isomorphic inner product spaces) *Two inner product spaces are isomorphic if and only if there is an inner product isomorphism between them.* ★

6.3 Orthogonality

We shall now introduce the concept of orthogonality which is the key to exploiting the rich structure of Hilbert spaces.

Definition 6.3.1 (Orthogonal vectors/sets, orthonormality) *Let V be a Hilbert space with inner product $\langle\cdot,\,\cdot\rangle$ and induced norm $\|\cdot\|$.*

1. *Let x and y be elements of V. x and y are said to be orthogonal to each other if and only if their inner product is zero, i.e. $\langle x,\, y\rangle = 0$.*
2. *Let x be an element of V and S be a subset of V. x is said to be orthogonal to the set S if and only if x is orthogonal to every element in S, i.e, $\langle x,\, y\rangle = 0$ for all $y \in S$.*
3. *Let S be a subset of V. S is said to be an orthogonal set if and only if each $x \in S$ is orthogonal to the set $S \setminus \{x\}$, i.e, $\langle x,\, y\rangle = 0$ for all $x, y \in S$, $x \neq y$.*
4. *A vector x in V is said to be normal if and only if $\|x\| = 1$.*
5. *A set S is orthonormal if and only if it is orthogonal and every element in it is normal.* $\star$

These definitions generalize the notion of a vector being perpendicular to another or a subset. It is important to recognize that orthogonality depends upon the inner product being used. Elements that are orthogonal to each other in some inner product may not be so in another inner product. Some examples will clarify these points.

Example 6.3.1 (Orthogonal vectors in $\mathbb{R}^n$) *Consider $\mathbb{R}^n$ as a Hilbert space with the inner product:*

$$\langle x,\, y\rangle = y^T x$$

which induces the Euclidean norm. Recall the standard basis introduced in Chapter 2:

$$\mathcal{B} = \{e_1, e_2, \cdots, e_n\} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \cdots, \begin{bmatrix} 0 \\ 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} \right\}$$

It is easily verified that

$$\langle e_i,\, e_j\rangle = e_j^T e_i = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

So, e_i is orthogonal to e_j whenever $i \neq j$. Every element of this basis is normal. Finally, this basis is an orthonormal set.

Now, suppose that we change the inner product from the standard inner product to the following weighted inner product:

$$\langle x,\, y\rangle_W = y^T W x$$

where $W > 0$ is a given non-diagonal matrix. $\mathbb{R}^n$ with this inner product is a Hilbert space. But, the set $\mathcal{B}$, which is still a basis for $\mathbb{R}^n$, is not necessarily orthonormal. This is because

$$\langle e_i,\, e_j\rangle_W = e_j^T W e_i = w_{ji}$$

which is not necessarily 0 or 1. $\triangle$

Example 6.3.2 (An orthogonal set in $\mathcal{L}_2([0,1])$) *Recall the Lebesgue space $\mathcal{L}_2([0,1])$ of square integrable functions defined on $[0,1]$ with inner product:*

$$\langle f,\, g\rangle = \int_0^1 f(t)g(t)dt$$

which is an infinite dimensional Hilbert space. Consider the following collection of functions in $\mathcal{L}_2([0,1])$:

$$\{\cos(2\pi kt)\}_{k=1}^{\infty}$$

We have:

$$\begin{aligned}
\langle \cos(2\pi kt),\, \cos(2\pi lt)\rangle &= \int_0^1 \cos(2\pi kt)\cos(2\pi lt)dt \\
&= \frac{1}{2}\int_0^1 \left(\cos\left(2\pi(k-l)t\right) + \cos\left(2\pi(k+l)t\right)\right)dt \\
&= \frac{1}{2}\begin{cases} \int_0^1 \left(1+\cos\left(4\pi kt\right)\right)dt & \text{if } k=l \\ \int_0^1 \left(\cos\left(2\pi(k-l)t\right) + \cos\left(2\pi(k+l)t\right)\right)dt & \text{otherwise} \end{cases} \\
&= \frac{1}{2}\begin{cases} 1 & \text{if } k=l \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

which shows that the set is orthogonal. △

Orthogonal sets have many nice properties. For example, if $\{x_k\}_{k=1}^n$ is an orthonormal set in the Hilbert space $\mathbb{R}^n$ with standard inner product, then the matrix

$$X = [\, x_1 \quad x_2 \quad \cdots \quad x_n \,]$$

is an orthonormal matrix (i.e., $X^TX = XX^T = I$). This is because the elements of the matrix X^TX are inner products of x_i and x_j which are either zero or one. The matrix X^TX is an example of a *Gram matrix* which we consider below. A further observation is that orthogonal sets are linearly independent.

Definition 6.3.2 (Gram matrix) *Let V be a Hilbert space with inner product $\langle\cdot,\,\cdot\rangle$ and $X = \{x_k\}_{k=1}^m$ be a collection of elements in V. The Gram matrix associated with the collection of elements is:*

$$G(X) = [\langle x_i,\, x_j\rangle]_{i,j=1}^m$$

and has size $m \times m$. ⋆

Proposition 6.3.1 (Testing for orthogonality) *Let V be a Hilbert space with inner product $\langle\cdot,\,\cdot\rangle$ and $\{x_k\}_{k=1}^{\infty}$ be a collection of elements in V. The following statements are true.*

1. *$\{x_k\}_{k=1}^{\infty}$ is an orthogonal set if and only if the Gram matrix $G(Y)$ associated with any finite subcollection $Y = \{y_i\}_{i=1}^m$ of $\{x_k\}_{k=1}^{\infty}$ is diagonal and invertible.*

2. *$\{x_k\}_{k=1}^{\infty}$ is an orthonormal set if and only if the Gram matrix $G(Y)$ associated with any finite sub-collection $Y = \{y_i\}_{i=1}^{m}$ of $\{x_k\}_{k=1}^{\infty}$ is the identity matrix.* ■

As a consequence, we have the following Pythagorus theorem.

Proposition 6.3.2 (Pythogorus theorem) *Let $\{x_k\}_{k=1}^{n}$ be an orthogonal set in the Hilbert space V. Then,*

$$\|x_1 + x_2 + x_3 + \cdots + x_n\|^2 = \|x_1\|^2 + \|x_2\|^2 + \|x_3\|^2 + \cdots + \|x_n\|^2$$

that is, the square of the norm of a sum of orthogonal vectors is the sum of the squares of the norms of the vectors. ■

6.4 Projection theorem

Many engineering problems are solved using the projection theorem. It is one of the most fundamental results in the Hilbert space theory. We state the theorem in two different ways and present some applications in this section.

Definition 6.4.1 (Orthogonal complement) *Let V be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and S be a non-empty subset of V. The orthogonal complement of S, denoted by $S^{\perp}$, is the set of all elements of V that are orthogonal to S, that is,*

$$S^{\perp} = \{x \in V : \langle x, y \rangle = 0 \text{ for all } y \in S\}$$

Put another way, x is in the orthogonal complement of S if and only if x is orthogonal to S. ★

Example 6.4.1 *Consider $\mathbb{R}^2$ with the standard inner product.*

1. *Let $S = \{0\}$. We claim that $S^{\perp} = \mathbb{R}^2$. This is easily seen as follows:*

$$\begin{aligned} S^{\perp} &= \left\{x \in \mathbb{R}^2 : \langle x, y \rangle = 0 \text{ for all } y \in S\right\} \\ &= \left\{x \in \mathbb{R}^2 : \langle x, 0 \rangle = 0\right\} = \mathbb{R}^2 \end{aligned}$$

2. *Now, suppose that*

$$S = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}$$

Then,

$$\begin{aligned} S^{\perp} &= \left\{x \in \mathbb{R}^2 : \langle x, y \rangle = 0 \text{ for all } y \in S\right\} \\ &= \left\{ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2 : \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0 \right\} \\ &= \left\{ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2 : x_1 + x_2 = 0 \right\} = \mathbf{Span}\left(\begin{bmatrix} 1 \\ -1 \end{bmatrix} \right) \end{aligned}$$

In these examples, eventhough S is not a subspace, $S^\perp$ is. It should also be noted that S and the span of S have the same orthogonal complement. △

Proposition 6.4.1 (Property of orthogonal complement) *Let V be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and S be a non-empty subset of V. The following statements are true.*

1. *$S^\perp$ is a closed subspace of V.*
2. *$S^\perp$ is the orthogonal complement of the span of S.* ■

It should be noted that a subspace of a Hilbert space is not necessarily closed. But, finite dimensional subspaces are closed. The first statement of the above proposition states that orthogonal complement of a set is a closed subset, always. According to the second statement, the orthogonal complement of a set is also the orthogonal complement of the span of the set. We are now ready to state the projection theorem.

Theorem 6.4.1 (Projection theorem - Abstract version) *Let V be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and S be a closed subspace of V. Then,*

$$\begin{aligned} V &= S \dot{\oplus} S^\perp \\ &= \left\{ z : \text{ there exist unique } x \in S \text{ and } y \in S^\perp \text{ such that } z = x + y \right\} \end{aligned}$$

that is, the Hilbert space is the direct orthogonal sum of S and its orthogonal complement. ■

This is perhaps the most important theorem in engineering applications. According to the theorem, given a closed subspace of a Hilbert space, we can decompose the Hilbert space into two components that are orthogonal to each other.

Example 6.4.2 (Application to $\mathbb{R}^2$) *Consider the Hilbert space $\mathbb{R}^2$ with the standard inner product $\langle \cdot, \cdot \rangle$. Let S be given by:*

$$S = \mathbf{Span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)$$

Then,

$$S^\perp = \mathbf{Span}\left(\begin{bmatrix} 1 \\ -1 \end{bmatrix} \right)$$

and by the projection theorem

$$\mathbb{R}^2 = \mathbf{Span}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) \dot{\oplus} \mathbf{Span}\left(\begin{bmatrix} 1 \\ -1 \end{bmatrix} \right)$$

This means that every $x \in \mathbb{R}^2$ can be written as the unique sum of a vector y in S and a vector z in $S^\perp$. Moreover, y and z are orthogonal to each other. △

We shall now state a more concrete version of the projection theorem. For this, the concept of an orthogonal projection is needed.

Definition 6.4.2 (Orthogonal projection) *Let V be a Hilbert space and S be a closed subspace of V. According to the projection theorem, each x in V can be written uniquely as:*

$$x = y + z$$

where $y \in S$ and $z \in S^{\perp}$.

1. *The component of x in S, namely y in the above decomposition, is called the orthogonal projection of x onto S.*
2. *The component of x in $S^{\perp}$, namely z in the above decomposition, is called the orthogonal projection of x onto $S^{\perp}$.*
3. *The map $\mathcal{P}_S : V \to V$ that takes x to y is called the orthogonal projection onto S.*
4. *The map $\mathcal{P}_{S^{\perp}} : V \to V$ that takes x to z is called the orthogonal projection onto $S^{\perp}$.* $\star$

Example 6.4.3 (Example 6.4.2 continued) *Note that any $x \in \mathbb{R}^2$ can be written as:*

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{2}(x_1 + x_2) \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \frac{1}{2}(x_1 - x_2) \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

The first term on the right hand side is the orthogonal projection of x onto S, and the second term is the orthogonal projection of x onto $S^{\perp}$.

Now, to compute the orthogonal projector $\mathcal{P}_S$, note that

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \overset{\mathcal{P}_S}{\to} \frac{1}{2}(x_1 + x_2) \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

under the projection operator. We can write this as:

$$\frac{1}{2}(x_1 + x_2) \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

so that

$$\mathcal{P}_S = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Similarly, we can compute $\mathcal{P}_{S^{\perp}}$. $\triangle$

Theorem 6.4.2 (Projection theorem - Concrete version) *Let V be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, S be a closed subspace of V and $x \in V$. Consider the minimization problem:*

$$\inf_{y \in S} \|x - y\|$$

that is, the problem of finding $y \in S$ that is closest to x. The following statements are true.

1. *The infimal value of the optimization problem is given by* $\|x - \mathcal{P}_S(x)\|$.

2. *There exists a unique* y *in* S *that achieves the infimal value. This unique solution is given by* $y = \mathcal{P}_S(x)$.

3. *The unique solution* $\mathcal{P}_S(x)$ *is orthogonal to the error* $x - \mathcal{P}_S(x)$*, that is,*

$$\langle \mathcal{P}_S(x),\, x - \mathcal{P}_S(x) \rangle = 0$$

Here, $\mathcal{P}_S(x)$ *is the orthogonal projection of* x *onto* S. ■

Chapter 7

Equilibrium Point and Linearization

Nonlinear systems are very difficult to analyze and there are very few useful truly nonlinear techniques. Fortunately, the interest in many applications is the analysis of systems about specific operating points. It is possible *sometimes* to approximate the nonlinear system in the vicinity of an operating point with a linear system and, more importantly, deduce properties of the nonlinear system from those of its linear approximation. The approximation process is called linearization and is the subject of this chapter. We may also be interested in transitioning the system from an operating point to another. Linearization is still useful, but it leads to more complicated time-varying linear systems.

7.1 Equilibrium point

7.1.1 Systems without inputs and outputs

Consider the nonlinear time-invariant continuous-time system:

$$\dot{x} = f(x) \tag{7.1}$$

where $x(t) \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}^n$ is a smooth function (meaning infinitely differentiable). The discrete-time analog is:

$$x(k+1) = f(x(k)) \tag{7.2}$$

where $x(k) \in \mathbb{R}^n$ and, as before, f is a smooth function from $\mathbb{R}^n$ into $\mathbb{R}^n$.

Definition 7.1.1 (Equilibrium point) *$x_0 \in \mathbb{R}^n$ is an equilibrium point of the continuous-time system (7.1) if and only if $f(x_0) = 0$. $x_0 \in \mathbb{R}^n$ is an equilibrium point of the discrete-time system (7.2) if and only if $f(x_0) = x_0$.* $\star$

Equilibrium points are also known as *zeros of f* (in the continuous-time case) and *fixed points*. If x_0 is an equilibrium point of the continuous-time system (7.1), then the differential equation (7.1) has a solution for

all time $t \geq 0$ starting at the initial condition $x(t=0) = x_0$. Indeed, one such solution is $x(t) = x_0$ for all $t \geq 0$. We shall refer to this solution as the *equilibrium solution*. Keep in mind that an equilibrium point is a point in $\mathbb{R}^n$, whereas an equilibrium solution is a function of time that happens to be constant. Thus, a continuous-time system starting at an equilibrium point stays there for all time. It is easy to show that if a continuous-time system enters an equilibrium state at time t_0, then it stays there for all time $t \geq t_0$. These statements hold for discrete-time systems also.

Example 7.1.1 *Consider the system*

$$\dot{x} = 0$$

where $x(t) \in \mathbb{R}^n$ (state space). Every point in $\mathbb{R}^n$ is an equilibrium point because the right hand side of the above equation is zero no matter which point in the state space is considered. On the other hand, the system

$$\dot{x} = x^2 + 1,$$

where $x(t) \in \mathbb{R}$, has no equilibrium point because the points where the right hand side becomes zero, i.e. $x^2 + 1 = 0 \Rightarrow x = \pm j$, are not in its state space. △

Example 7.1.2 *Consider the system*

$$\dot{x} = \sin(x)$$

where $x(t) \in \mathbb{R}$. To find the equilibrium points, we look for those points in the state space where the right hand side is zero. That is, find all the solutions in $\mathbb{R}$ of:

$$\sin(x) = 0$$

The equilibrium points are clearly given by: $x = \cdots, -2\pi, -\pi, 0, \pi, 2\pi, \cdots$. △

An equilibrium point (of a continuous-time or discrete-time system) is *isolated* if it has an open neighborhood that contains no other equilibrium point. Nonlinear systems can have isolated equilibrium points, a dense set of equilibrium points, etc (just about anything you can imagine).

Example 7.1.3 *The system in Example 7.1.2 has only isolated equilibrium points; the first system in Example 7.1.1 has a continuum of equilibrium points.* △

7.1.2 Systems with inputs

Consider the nonlinear time-invariant continuous-time system:

$$\dot{x} = f(x, u) \tag{7.3}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$ and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is a smooth function. Its discrete-time analog is:

$$x(k+1) = f(x(k), u(k)) \tag{7.4}$$

where as before k is the discrete-time index.

Definition 7.1.2 (Equilibrium point) *A pair* $(x_0, u_0) \in \mathbb{R}^n \times \mathbb{R}^m$ *is an equilibrium point of the continuous-time system (7.3) if* $f(x_0, u_0) = 0$. *A pair* $(x_0, u_0) \in \mathbb{R}^n \times \mathbb{R}^m$ *is an equilibrium point of the discrete-time system (7.4) if* $f(x_0, u_0) = x_0$. ⋆

Note that the definition involves a state vector x_0 and a control input vector u_0. This type of equilibrium point is also known as a *trim point* and, in some cases, an *operating point*.

Example 7.1.4 *Consider*

$$\dot{x} = x^2 + u$$

where $x(t)$ *and* $u(t)$ *are real numbers. To find the trim points, we solve for all real solutions of:*

$$x^2 + u = 0$$

and obtain the set:

$$\{(\sqrt{v}, -v), (-\sqrt{v}, -v) : v \in [0, \infty)\}$$

as the set of all solutions. Every point in the above set is an equilibrium point. The left hand side plot in figure 7.1 shows part of this set.

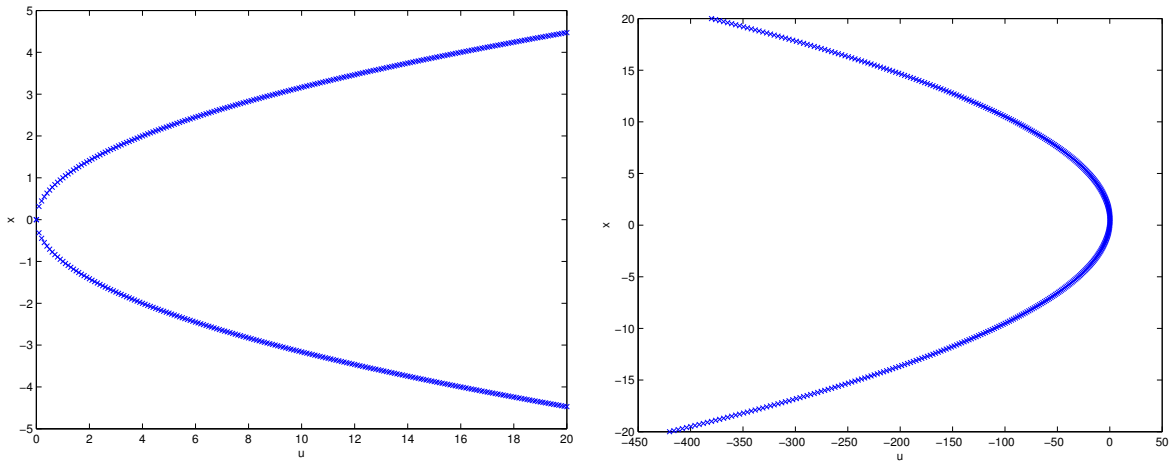

Figure 7.1: Trim point curves for continuous-time system (left) and discrete-time system (right)

On the other hand, the equilibrium points of the discrete-time system:

$$x(k+1) = x(k)^2 + u(k),$$

where $x(k)$ *and* $u(k)$ *are real numbers, are given by real solutions of:*

$$x = x^2 + u$$

The right hand side plot in figure 7.1 shows part of the curve defined by the above equation. △

7.2 Linearization about an equilibrium point

Let x be a vector in $\mathbb{R}^n$ and denote by x_i the ith component of x, i.e.,

$$x = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix}$$

Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a smooth function. For each $x \in \mathbb{R}^n$, the value of f at x, namely $f(x)$, is a vector in $\mathbb{R}^n$ which can be written in component form as:

$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ . \\ . \\ . \\ f_n(x) \end{bmatrix}$$

Here, f_i is the ith component of f.

Example 7.2.1 *Define $f : \mathbb{R}^2 \to \mathbb{R}^2$ as follows:*

$$f(x) = \begin{bmatrix} x_1^2 + x_2^2 \\ x_2 + \sin(x_2) \end{bmatrix}$$

where x_1 and x_2 are the components of x. f is a function of two variables (x_1, x_2) and takes values in $\mathbb{R}^2$. Moreover,

$$f_1(x) = x_1^2 + x_2^2, \quad f_2(x) = x_2 + \sin(x_2)$$

are the components of f. △

Definition 7.2.1 (Jacobian) *The* Jacobian of f at x *is the $n \times n$ matrix defined as:*

$$\frac{\partial f}{\partial x} = \begin{bmatrix} (\partial f_1/\partial x_1) & (\partial f_1/\partial x_2) & \cdot & \cdot & \cdot & (\partial f_1/\partial x_n) \\ (\partial f_2/\partial x_1) & (\partial f_2/\partial x_2) & \cdot & \cdot & \cdot & (\partial f_2/\partial x_n) \\ . & . & . & . & . & . \\ (\partial f_n/\partial x_1) & (\partial f_n/\partial x_2) & \cdot & \cdot & \cdot & (\partial f_n/\partial x_n) \end{bmatrix}$$

where $(\partial f_i/\partial x_j)$ is the partial derivative of the ith component of f with respect to the jth component of x. ⋆

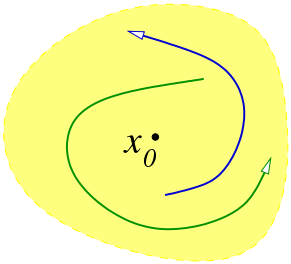

Figure 7.2: Trajectories emanating from initial conditions in the vicinity of an equilibrium point

The Jacobian of f given in Example 7.2.1 can be computed easily as:

$$\begin{bmatrix} 2x_1 & 2x_2 \\ 0 & 1+\cos(x_2) \end{bmatrix}$$

An important point to note is that the Jacobian is a matrix-valued function. It becomes a matrix only when it is evaluated at a particular x. We use the notation:

$$\frac{\partial f}{\partial x}\Big|_{x_0}$$

for *the Jacobian of f evaluated at $x = x_0$*. This is a $n \times n$ constant matrix. Although our definition applies to functions from $\mathbb{R}^n$ into $\mathbb{R}^n$, the extension to functions mapping $\mathbb{R}^n$ into $\mathbb{R}^m$ is clear.

7.2.1 Systems without inputs

Consider the continuous-time system given in (7.1). Since f is smooth, it has a Taylor's series expansion about x_0:

$$f(x) = f(x_0) + \frac{\partial f}{\partial x}\Big|_{x_0} (x - x_0) + \text{ higher order terms}$$

which is valid for all x in an open neighborhood of x_0. Note that the coefficient of $(x - x_0)$ in the Taylor's series expansion is the Jacobian of f evaluated at $x = x_0$. Let us denote this matrix by A. If x_0 is an equilibrium point of the continuous-time system, then $f(x_0) = 0$ and

$$f(x) \approx \frac{\partial f}{\partial x}\Big|_{x_0} (x - x_0) = A(x - x_0)$$

which is accurate up to first order. That is, in the vicinity of an equilibrium point, the (possibly) nonlinear differential equation (7.1) can be written as:

$$\dot{x} = A(x - x_0) + \text{ small terms} \tag{7.5}$$

Now, let us think in terms of solutions of the differential equation (7.1). Since x_0 is an equilibrium point by assumption, it is an equilibrium solution. Pick an initial condition x_{ic} that is close to x_0. It can be proved

using smoothness of f that there is a time interval $[0, T)$ over which the nonlinear system has a solution starting from x_{ic}. Along this solution, the expression:

$$\dot{x} - A(x - x_0) \approx 0$$

holds for as long as the solution stays in the vicinity of x_0 where (7.5) is valid. In fact, there is a time $0 < T_1 \leq T$ such that the approximate condition holds for all time in the interval $[0, T_1)$. We can write the above expression as:

$$\frac{d}{dt}(x - x_0) - A(x - x_0) \approx 0 \tag{7.6}$$

since x_0 is a constant.

Consider the linear time-invariant (LTI) system:

$$\dot{z} = Az$$

with initial condition $z(0) = x_{ic} - x_0$. Define

$$\hat{x} = x_0 + z$$

We claim that $\hat{x}$ is close to x. Indeed,

$$\dot{x} - \dot{\hat{x}} = f(x) - \frac{d}{dt}(x_0 + z) = f(x) - \dot{z} = f(x) - Az = f(x) - A(x - x_0) \approx 0$$

which along with $x(0) - \hat{x}(0) = 0$ implies that x and $\hat{x}$ are close to each other in a sufficiently small time interval. In summary, the solution z of the LTI system is such that $x_0 + z$ is close to the solution of the nonlinear system for at least a small time interval. This is true of any initial condition x_{ic} that is close to the equilibrium point x_0. The equation $\dot{z} = Az$ is the linearization of the nonlinear system about the equilibrium point x_0. Similar argument can be made for the discrete-time system (7.2) to arrive at the linear difference equation $z_{k+1} = Az_k$.

Definition 7.2.2 (Linearization - no input case) *The linearization of a continuous-time (resp. discrete-time) nonlinear system (7.1) about the equilibrium point x_0 (resp. (7.2) about x_0) is given by*

$$\begin{aligned} \textit{Continuous-Time:} \quad & \dot{z} = Az \\ \textit{Discrete-Time:} \quad & z_{k+1} = Az_k \end{aligned}$$

where A is the Jacobian of f evaluated at x_0. ⋆

Linearization of a nonlinear system is a linear dynamical system. When the linearization is performed about an equilibrium point of a time-invariant nonlinear system, then the linearization is a linear time-invariant (LTI) dynamical system.

Example 7.2.2 *The equilibrium points of the system:*

$$\dot{x} = \sin(x),$$

where $x(t) \in \mathbb{R}$, were determined earlier to be:

$$\{\cdots, -2\pi, -\pi, 0, \pi, 2\pi, \cdots\} = \{k\pi\}_{k=-\infty}^{\infty}$$

The Jacobian of "f" evaluated at a generic equilibrium point $x_0 = k\pi$ is:

$$\frac{d}{dx}\sin(x)|_{x=k\pi} = \cos(x)|_{x=k\pi} = (-1)^{|k|}$$

So, the linearization at $x_0 = k\pi$ is:

$$\dot{x} = \begin{cases} x & \text{if } |k| \text{ is zero or even} \\ -x & \text{otherwise} \end{cases}$$

for any integer k. △

7.2.2 Systems with inputs

Consider the continuous-time system given in (7.3). The Taylor's series expansion of f about (x_0, u_0) is:

$$f(x,u) = f(x_0,u_0) + \frac{\partial f}{\partial x}\Big|_{(x_0,u_0)}(x - x_0) + \frac{\partial f}{\partial u}\Big|_{(x_0,u_0)}(u - u_0) + \text{ higher order terms}$$

which is valid for all (x, u) in an open neighborhood of (x_0, u_0). The symbol

$$\frac{\partial f}{\partial u}$$

stands for the Jacobian of f with respect to u and has a definition similar to the Jacobian given in Definition 7.2.1. Let us denote these two Jacobians by A and B respectively. Proceeding as before, we arrive at the following definition of linearization.

Definition 7.2.3 (Linearization - with inputs) *The linearization of the continuous-time (resp. discrete-time) system with inputs (7.3) about the equilibrium point (x_0, u_0) (resp. (7.4) about (x_0, u_0)) is given by:*

$$\begin{aligned} \textit{Continuous-Time:} \quad & \dot{z} = Az + Bv \\ \textit{Discrete-Time:} \quad & z_{k+1} = Az_k + Bv_k \end{aligned}$$

where A and B are the Jacobians of f with respect to x and u evaluated at (x_0, u_0). ⋆

7.2.3 Systems with inputs and outputs

Consider the continuous-time system given by (7.3) along with the output equation:

$$y = h(x, u)$$

Let (x_0, u_0) be an equilibrium point of this system. Assume that $h(x_0, u_0) = 0$. Then, the linearization about (x_0, u_0) is given by:

$$\dot{z} = Az + Bv$$
$$\hat{y} = Cz + Dv$$

where

$$A = \frac{\partial f}{\partial x}\Big|_{(x_0,u_0)}, \quad B = \frac{\partial f}{\partial u}\Big|_{(x_0,u_0)}, \quad C = \frac{\partial h}{\partial x}\Big|_{(x_0,u_0)} \quad \text{and} \quad D = \frac{\partial h}{\partial u}\Big|_{(x_0,u_0)}$$

Example 7.2.3 *It is easy to see by substitution that* $(x = 0, u = 0)$ *is an equilibrium point of*

$$\dot{x} = -x^4 \sin(x) + \sin u$$
$$y = x + u \cos u$$

To linearize about this equilibrium point, we first calculate and evaluate the Jacobians:

$$A = \frac{\partial f}{\partial x}\Big|_{(x_0,u_0)} = \frac{d}{dx}\left(-x^4 \sin(x) + \sin u\right)\Big|_{(0,0)} = \left(-4x^3 \sin(x) - x^4 \cos(x)\right)\Big|_{(0,0)} = 0$$

$$B = \frac{\partial f}{\partial u}\Big|_{(x_0,u_0)} = \frac{d}{du}\left(-x^4 \sin(x) + \sin u\right)\Big|_{(0,0)} = (\cos(u))\Big|_{(0,0)} = 1$$

$$C = \frac{\partial h}{\partial x}\Big|_{(x_0,u_0)} = \frac{d}{dx}\left(x + u\cos(u)\right)\Big|_{(0,0)} = (1)\Big|_{(0,0)} = 1$$

$$D = \frac{\partial h}{\partial u}\Big|_{(x_0,u_0)} = \frac{d}{du}\left(x + u\cos(u)\right)\Big|_{(0,0)} = (\cos(u) - u\sin(u))\Big|_{(0,0)} = 1$$

Then, we form:

$$\dot{z} = Az + Bv = v$$
$$\hat{y} = Cz + Dv = z + v$$

which is the linearization about $(0, 0)$. △

Chapter 8

LTI System Behavior

A continuous-time LTI system has the form:

$$\dot{x} = Ax + Bu, \qquad x(0) = x_0 \tag{8.1a}$$

$$y = Cx + Du \tag{8.1b}$$

where $x(t) \in \mathbb{R}^n$ is the state, $u(t) \in \mathbb{R}^m$ is the input and $y(t) \in \mathbb{R}^p$ is the output. The state space matrices (A, B, C, D) are constant real matrices of appropriate dimensions and x_0 is the initial condition. *A solution of the differential equation (8.1a)* emanating from x_0 in response to the input u is a function $x : [0, \infty) \to \mathbb{R}^n$ such that

$$x(0) = x_0 \quad \text{and} \quad \dot{x} = Ax + Bu$$

i.e., it must satisfy the initial condition and the differential equation (8.1a). An explicit formula for the solution is derived in this chapter. Once the formula for x is known, we can easily obtain an expression for the output y by simply substituting the formula for x into (8.1b). The discrete-time LTI system:

$$x_{k+1} = Ax_k + Bu_k, \qquad x(0) = x_0 \tag{8.2a}$$

$$y_k = Cx_k + Du_k \tag{8.2b}$$

will also be studied; it turns out to be much easier than the continuous-time system.

8.1 Solution of continuous-time LTI equations

The solution is given in three steps. First, we consider the case without inputs:

$$\dot{x} = Ax, \quad x(0) = x_0 \tag{8.3}$$

that is, $u = 0$ in equation (8.1a). The system responds in this case to the initial condition and, hence, the solution is called *initial condition* or *unforced* or *zero input* response. In the second step, we set the initial

condition to zero and consider the effect of the input u. The solution in this case is called *input* or *forced* or *zero initial condition* response. Finally, the general case with both initial conditions and inputs is considered. We are able to do this because of a fundamental consequence of linearity, namely the *superposition principle*. Roughly speaking, the principle states that the total effect of a number of causes is the sum of the effects of the individual causes. So, the solution in the general case is simply the sum of initial condition response and forced response.

Theorem 8.1.1 (Initial condition response) *The solution x of (8.3) starting at the initial condition x_0 is given by:*

$$x(t) = e^{At}x_0$$

for all $t \geq 0$. ■

Definition and properties of the exponential function e^A can be found in Chapter 4. It plays a key role in describing the initial condition response and, as we shall soon see, forced response. Some features of initial condition response that readily follow from linearity and properties of the exponential function are listed below.

Theorem 8.1.2 (Properties of initial condition response) *The following statements are true.*

1. *Suppose that x and y are initial condition responses starting from x_0 and y_0 respectively at time $t = 0$. Let α and β be two real numbers. Then, $\alpha x + \beta y$ is the initial condition response starting from $\alpha x_0 + \beta y_0$ at time $t = 0$.*
2. *Suppose that (λ, v) is an eigenvalue-eigenvector pair of A. Then, for any initial condition in the eigen-subspace spanned by v, the corresponding initial condition response stays in the eigen-subspace spanned by v for all time. That is,*

$$x(0) = \alpha v \text{ for some real number } \alpha \Rightarrow x(t) = \alpha e^{\lambda t} v \text{ for all } t \geq 0$$

3. *Suppose that $\mathcal{E}$ is an eigen-subspace of A. Then, for any initial condition $x_0 \in \mathcal{E}$, the corresponding response x is such that $x(t) \in \mathcal{E}$ for all $t \geq 0$.* ■

The first statement follows from linearity. The last two statements are very important. They say that if initial conditions are chosen to lie in eigen-subspaces of A, then the response never leaves the subspace. In mathematical terms, we say that *eigen-subspaces of A are flow-invariant*. The unforced system (8.3) has other important system-theoretic properties that are algebraic or group-theoretic in nature. To discuss these, we make the following important definition.

Definition 8.1.1 (Transition matrix) *The transition matrix of the LTI system (8.1) is defined as:*

$$\phi(t, s) = e^{A(t-s)}$$

for all t, s. ★

The transition matrix is a matrix-valued function of two time indices, t and s. When $t \geq s$, we can think of $t - s$ as the *elapsed time*. Clearly, the initial condition response given in Theorem 8.1.1 can be written as $x(t) = \phi(t, 0)x_0$. This equation can be interpreted as follows. Note that $s = 0$ and, hence, the elapsed time is t. So, if the system is at the state x_0, then after an elapsed time of t, the system will be at the state $\phi(t, 0)x_0$. The next theorem describes properties of ϕ and provides an intuitive explanation of what it means to be a transition matrix of a system.

Theorem 8.1.3 (Properties of transition matrix) *Let ϕ denote the transition matrix of the system (8.3) with no inputs. The following statements are true.*

1. *$\phi(t, t) = I$ for any t, i.e., the state of a system is the same after $t - t = 0$ time has elapsed.*
2. *$\phi(s, t) = \phi(t, s)^{-1}$ for any t and s. Thus, if $x(t)$ and $x(s)$ are two points on the solution x at times t and s, then they are related as follows:*
$$x(s) = \phi(s, t)x(t) = e^{A(s-t)}x(t) = e^{-A(t-s)}x(t) = \phi(t, s)^{-1}x(t)$$
3. *$\phi(u, t) = \phi(u, s)\phi(s, t)$ for any t, s and u, .i.e., transitioning from t to u is same as transitioning from t to s and then from s to u. Thus,*
$$x(u) = \phi(u, t)x(t) = \phi(u, s)\left(\phi(s, t)x(t)\right) = \phi(u, s)x(s)$$
for any three points $x(t)$, $x(s)$ and $x(u)$ on the solution.
4. *$\phi(t, s)$ satisfies:*
$$\frac{d}{dt}\phi(t, s) = A\phi(t, s)$$
for any t and s. ■

An important property that is implicit in the above theorem is that the transition matrix is invertible for all values of its arguments. It is a consequence of the definition of transition matrix; from a system-theoretic point, it means that a starting point and an elapsed time completely defines the end point and vice-versa.

Example 8.1.1 *Let*
$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$
which is in Jordan form. Hence, by Theorem 4.2.4,
$$e^{At} = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}$$
for all t. Therefore, the transition matrix of (8.3) with this particular A is:
$$\phi(t, s) = \begin{bmatrix} 1 & t - s \\ 0 & 1 \end{bmatrix}$$
for all t, s. △

Let us now consider the case with input u and zero initial condition. We have the following theorem.

Theorem 8.1.4 (Forced response) *The solution x of (8.1a) with initial condition $x_0 = 0$ is given by:*

$$x(t) = \int_0^t e^{A(t-\tau)} B u(\tau) d\tau$$

for all $t \geq 0$. ■

The integral appearing above is an example of a *convolution integral* which is a fundamental quantity in systems theory. In terms of the transition matrix ϕ, the forced solution is:

$$x(t) = \int_0^t \phi(t, \tau) u(\tau) d\tau$$

It is generally difficult to evaluate these convolutions. The Laplace transform can be used in some special cases. But, numerical integration methods are usually applied on the differential equation (8.1a) to get the solution. The explicit formula for x in Theorem 8.1.4 has important theoretical applications and will be often used in subsequent chapters. Proof of the theorem is easy and involves applying the Leibniz rule:

Theorem 8.1.5 (Calculus and Leibniz rule) *Suppose that ϕ is a matrix valued differentiable function of two scalar variables t and τ. Then,*

$$\frac{d}{dt} \int_{f(t)}^{g(t)} \phi(t, \tau) d\tau = \phi(t, g(t)) \frac{dg}{dt}(t) - \phi(t, f(t)) \frac{df}{dt}(t) + \int_{f(t)}^{g(t)} \frac{\partial}{\partial t} \phi(t, \tau) d\tau$$

for any differentiable functions f and g. ■

We complete the solution of the system (8.1) below. As mentioned, since the system is linear, the total solution is just the sum of initial condition response given by Theorem 8.1.1 and forced response given by Theorem 8.1.4.

Theorem 8.1.6 (Complete solution) *Consider the system (8.1). The state trajectory x starting at the initial condition x_0 and evolving under the input u is given by:*

$$x(t) = e^{At} x_0 + \int_0^t e^{A(t-\tau)} B u(\tau) d\tau$$

for all $t \geq 0$. The corresponding output trajectory y is given by:

$$y(t) = Cx(t) + Du(t) = C \left(e^{At} x_0 + \int_0^t e^{A(t-\tau)} B u(\tau) d\tau \right) + Du(t)$$

for all $t \geq 0$. ■

8.2 Solution of discrete-time LTI equations

The solution in the discrete time case is easy to obtain as the defining relation (8.2a) is a difference equation and not a differential equation. Clearly, when $k = 1$, the difference equation (8.2a) gives:

$$x_1 = Ax_0 + Bu_0$$

which is a formula for the state at time $k = 1$ in terms of the initial condition x_0 and the value u_0 of the input u at time $k = 0$. Let us set $k = 2$. Then, from the difference equation and the above formula for x_1, we get:

$$\begin{aligned} x_2 &= Ax_1 + Bu_1 \\ &= A(Ax_0 + Bu_0) + Bu_1 \\ &= A^2x_0 + ABu_0 + Bu_1 \end{aligned}$$

which is a formula for the state at time $k = 2$. Now, set $k = 3$ and proceed as before to get:

$$\begin{aligned} x_3 &= Ax_2 + Bu_2 \\ &= A\left(A^2x_0 + ABu_0 + Bu_1\right) + Bu_2 \\ &= A^3x_0 + A^2Bu_0 + ABu_1 + Bu_2 \\ &= A^3x_0 + \left(\sum_{i=0}^{2} A^{2-i}Bu_i\right) \end{aligned}$$

Repeating this procedure of applying the difference equation and eliminating variables using previously obtained formulas leads to the following theorem.

Theorem 8.2.1 (Complete solution for discrete-time system) *Consider the system (8.2). The state trajectory x starting at the initial condition x_0 and evolving under the input u is given by:*

$$x_k = A^kx_0 + \sum_{i=0}^{k-1} A^{k-1-i}Bu_i$$

for all $k \geq 0$. The corresponding output trajectory y is given by:

$$y_k = Cx_k + Du_k = C\left(A^kx_0 + \sum_{i=0}^{k-1} A^{k-1-i}Bu_i\right) + Du_k$$

for all $k \geq 0$. ■

As in the continuous-case, the state trajectory is the sum of an *initial condition response*

$$A^kx_0$$

and a *forced response*:

$$\sum_{i=0}^{k-1} A^{k-1-i}Bu_i$$

The forced response is now a sum, the discrete analog of an integral involved in the continuous-time case. This sum is an example of a *discrete convolution*. Another important point to note is that *the transition matrix in the discrete-time case* is given by:

$$\phi(k, m) = A^{k-m}$$

where k and m are any two time indices. It has all the properties stated in Theorem 8.1.3. Finally, for later use, let us write the formula for x in the following form:

$$\begin{aligned} x_k &= A^k x_0 + \sum_{i=0}^{k-1} A^{k-1-i} B u_i \\ &= A^k x_0 + \begin{bmatrix} B & AB & A^2B & \cdots & A^{k-1}B \end{bmatrix} \begin{bmatrix} u_{k-1} \\ u_{k-2} \\ \cdot \\ \cdot \\ u_1 \\ u_0 \end{bmatrix} \end{aligned}$$

The matrix

$$\begin{bmatrix} B & AB & A^2B & \cdots & A^{k-1}B \end{bmatrix}$$

will play an important role in later chapters.

Chapter 9

Lyapunov (Internal) Stability Notions

This chapter introduces several concepts of internal stability of state space models of dynamical systems. They originated in the work of A. M. Lyapunov and, hence, are refered to as Lyapunov stability. Nonlinear systems can also exhibit stable behaviors that fall outside Lyapunov theory. We shall not consider such behaviors.

9.1 Nonlinear time-invariant systems

Consider the continuous-time system:

$$\dot{x} = f(x), \qquad x(0) = x_0 \tag{9.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is smooth. We will focus exclusively on continuous-time systems as the definitions can be adapted to the discrete-time case. Throughout, x denotes the solution of (9.1) emanating from x_0 at $t = 0$ (if it exists). The value of this solution at a specific point in time t will be denoted by $x(t)$. We use $\|v\|$ to denote the Euclidean norm of a vector v.

Definition 9.1.1 (Boundedness) *A solution x of (9.1) is bounded if and only if there exists $\beta \geq 0$ such that*

$$\|x(t)\| \leq \beta \ \textit{for all} \ \ t \geq 0$$

A solution that is not bounded is called unbounded. ⋆

The definition means that the solution is contained in a ball in $\mathbb{R}^n$ of radius β centered at 0 for all time. The radius of the ball β could depend on the initial condition x_0.

Example 9.1.1 *The solution of $\dot{x} = 0$ starting from any initial condition is bounded because $x(t) = x_0$ for all $t \geq 0$* (estimate β in the definition). *All solutions, except the zero solution starting from $x_0 = 0$, of*

$\dot{x} = x$ are unbounded because $x(t) = e^t x_0$ which grows beyond any finite ball except when $x_0 = 0$. All solutions of $\dot{x} = -x^3$ are bounded. △

Example 9.1.2 *Consider the LTI systems:*

$$\dot{x}_1 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} x_1 \text{ and } \dot{x}_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x_2$$

Clearly,

$$x_1(t) = e^{\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} t} x_1(0) = x_1(0)$$

and

$$x_2(t) = e^{\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} t} x_2(0) = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} x_2(0)$$

x_1 is bounded for any initial condition; but x_2 is not always bounded. For example, with the initial condition:

$$x_2(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

we get:

$$x_2(t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{for all } t \geq 0$$

which is bounded; but with the initial condition:

$$x_2(0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

we get:

$$x_2(t) = \begin{bmatrix} 1 + 2t \\ 2 \end{bmatrix} \quad \text{for all } t \geq 0$$

which is unbounded since the first component grows linearly with time. △

Definition 9.1.2 (Local Stability (LS)) *Let x_e be an equilibrium point of the system given by (9.1). x_e is locally stable in the sense of Lyapunov if and only if for each $\epsilon > 0$ there exists $\delta > 0$ such that whenever the initial condition $x(0)$ satisfies*

$$\|x(0) - x_e\| < \delta,$$

the resulting solution x satisfies

$$\|x(t) - x_e\| < \epsilon \text{ for all } t \geq 0$$

An equilibrium point that is not locally stable is called unstable. ⋆

Clearly, if the initial condition is taken to be $x(0) = x_e$, then the corresponding solution is $x(t) = x_e$ and we have $\|x(t) - x_e\| = 0$ for all time. For stability, we need to know how the solutions starting from a neighborhood of the equilibrium point behave. The definition means the following. Suppose that the system is LS about x_e. Pick a ball centered at x_e of radius $\epsilon > 0$. Then, we can find a ball centered at x_e of radius $\delta > 0$ such that if the system is started from any point inside the δ-ball, the corresponding solution will not leave the ϵ-ball. There are a couple of things to note:

- The ϵ-ball is picked first. So, δ depends on ϵ.
- The definition says that some $\delta > 0$ must exist for every $\epsilon > 0$. However, we have the following. Suppose that for a certain $\epsilon_0 > 0$ there is a $\delta_0 > 0$ that satisfies the requirements in the definition. Then, there is an obvious choice for δ that works for any $\epsilon > \epsilon_0$.
- LS implies boundedness of solutions emanating from the vicinity of x_e. Boundedness of solutions does not imply LS (for nonlinear systems). A good example for this is a system exhibiting a limit cycle.

Example 9.1.3 *Every $v \in \mathbb{R}^n$ is a LS equilibrium point of $\dot{x} = 0$. 0 is an unstable equilibrium point of $\dot{x} = x$. 0 is a LS equilibrium point of $\dot{x} = -x^3$.* △

Example 9.1.4 (Bounded Solutions, Unstable Equilibrium Point) *Consider the Van der Pol oscillator*

$$\begin{aligned}\dot{x}_1 &= x_2\\ \dot{x}_2 &= (1 - x_1^2)x_2 - x_1\end{aligned}$$

All solutions are bounded; but the equilibrium point $(0, 0)$ is unstable. △

Definition 9.1.3 (Local Attractor) *An equilibrium point x_e of the system (9.1) is a local attractor if and only if there exists $\beta > 0$ such that whenever the initial condition $x(0)$ satisfies*

$$\|x(0) - x_e\| < \beta,$$

the resulting solution satisfies

$$\lim_{t\to\infty} x(t) = x_e$$

(i.e. converges asymptotically to the equilibrium point). ⋆

Definition 9.1.4 (Local Asymptotic Stability (LAS)) *Let x_e be an equilibrium point of the system given by (9.1). x_e is locally asymptotically stable in the sense of Lyapunov if and only if the following two conditions hold:*

1. *x_e is LS*

2. *x_e is a local attractor.* ■

The second condition in the definition means that all solutions starting from a neighborhood converge to the equilibrium point as time goes to ∞. In other words, a LAS equilibrium point is also an *attractor*. We might think that the converse is true, i.e. if all solutions starting from a neighborhood of the equilibrium point come arbitrarily close to it after some time (or even converge to as t goes to ∞) the system must be LAS. This however is false. There are unstable nonlinear systems whose solutions starting from a neighborhood converge to the equilibrium point. This is why in the definition of LAS we have the first condition.

LAS does not say anything about how fast we converge to x_e. The next definition involves rate of convergence.

Definition 9.1.5 (Local Exponential Stability (LES)) *Let x_e be an equilibrium point of the system given by (9.1). x_e is locally exponentially stable in the sense of Lyapunov if and only if the following two conditions hold:*

1. *x_e is LAS*
2. *There exist $\alpha > 0$ and $\gamma > 0$ such that for any convergent solution x starting at $x(0)$, we have*

$$\|x(t) - x_e\| \leq \alpha \|x(0)\| e^{-\gamma t}$$

for all $t \geq 0$. ■

In the definitions given above, the initial conditions were chosen from a neighborhood of the equilibrium point. Hence, these notions are *local* in nature. The next definitions are global.

Definition 9.1.6 (Global Asymptotic Stability (GAS)) *Let x_e be an equilibrium point of the system (9.1). x_e is globally asymptotically stable in the sense of Lyapunov if and only if the following two conditions hold:*

1. *x_e is LS*
2. *Any solution x satisfies*

$$\lim_{t \to \infty} x(t) = x_e$$

■

Compare this definition with LAS definition.

Example 9.1.5 *The system $\dot{x} = -x$ is GAS. The damped simple pendulum*

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\sin x_1 - x_2 \end{aligned}$$

is LAS but not GAS because it has more than one equilibrium point. In fact, a necessary condition for GAS is the existence of one and only one equilibrium point. ■

Definition 9.1.7 (Global Exponential Stability (GES)) *Let x_e be an equilibrium point of the system given by (9.1). x_e is globally exponentially stable (in the sense of Lyapunov) if and only if the following two conditions hold:*

1. *x_e is GAS*
2. *There exist $\alpha > 0$ and $\gamma > 0$ such that for any solution x starting at $x(0)$, we have*

$$\|x(t) - x_e\| \leq \alpha \|x(0)\| e^{-\gamma t}$$

for all $t \geq 0$. ■

9.1.1 Relations between stability notions

The different notions of stability defined above are related in the following way.

$$\begin{array}{ccccccc} \text{LES} & \Rightarrow & \text{LAS} & \Rightarrow & \text{LS} & \Rightarrow & \text{Locally Bounded} \\ \Uparrow & & \Uparrow & & \Uparrow & & \Uparrow \\ \text{GES} & \Rightarrow & \text{GAS} & \Rightarrow & \text{GS} & \Rightarrow & \text{Bounded} \end{array}$$

The general idea is that global properties imply local properties, and exponential convergence implies asymptotic convergence. The converse is not true in general. For (finite dimensional) linear systems, the converse is true. That is, local implies global, and asymptotic implies exponential. This is discussed in the next section.

9.2 LTI systems

When the system under consideration is LTI, the nonlinear differential equations of (9.1) become:

$$\dot{x} = Ax, \qquad x(0) = x_0 \tag{9.2}$$

where A is a constant matrix. Applying the results of Chapter 8, the solution x can be written as:

$$x(t) = e^{At} x_0$$

This explicit formula for the solution, which is unavailable in the general nonlinear case, can be used to great advantage in studying LTI stability. In fact, we shall use this formula and properties of the exponential function to derive several simplifications.

Boundedness of solutions and LS are different concepts for nonlinear systems. This is illustrated by the Van der Pol oscillator example 9.1.4. They are equivalent concepts for LTI systems.

Theorem 9.2.1 (LS vs. bounded solutions) *The following statements are true for the LTI system (9.2).*

1. 0 *is LS if and only if solution emanating from any initial condition is bounded.*
2. *An equilibrium point is LS if and only if solution emanating from any initial condition is bounded.*
3. *An equilibrium point is LS if and only if every equilibrium point is LS.* ■

Proof of this theorem relies on two properties of linear systems (i) scaling initial conditions simply scales the solution by the same amount, and (ii) if the solutions starting from a finite set of initial conditions are bounded, then so are the solutions starting from any linear combination of the initial conditions. These are properties not satisfied by general nonlinear systems. Statement 3 is a trivial consequence of Statements 1 and 2. We have stated it because it indicates that *for LTI systems, LS is a property of the system and is independent of the equilibrium point*.

The next result shows that local and global notions of stability are equivalent for LTI systems.

Theorem 9.2.2 (Local vs. global) *The following statements are true for the LTI system (9.2).*

1. 0 *is LAS if and only if* 0 *is GAS.*
2. 0 *is LAS if and only if* 0 *is LES.* ■

These statements suggest that there is no need to distinguish between LAS, GAS, LES and GES for LTI systems. They all mean the same thing. Proof again relies on linearity. For instance, to show GAS from LAS, we use scaling argument. The equivalence of LAS and LES is somewhat different and involves estimating the norm of the transition matrix.

The results cited above allow us to compile the following table for LTI systems.

$$\begin{array}{ccccccc} \text{LES} & \Leftrightarrow & \text{LAS} & \Rightarrow & \text{LS} & \Leftrightarrow & \text{Locally Bounded} \\ \Updownarrow & & \Updownarrow & & \Updownarrow & & \Updownarrow \\ \text{GES} & \Leftrightarrow & \text{GAS} & \Rightarrow & \text{GS} & \Leftrightarrow & \text{Bounded} \end{array}$$

9.3 Lyapunov's direct method

There are two ways to establish asymptotic stability of a nonlinear systems. The first method called *Lyapunov's direct method* is based on constructing an energy-like function known as the *Lyapunov function* and showing that energy decreases along the trajectories of the system. This idea originated from mechanical systems where the notions of kinetic and potential energies are well-established. In mechanical systems, the total energy which is the sum of kinetic and potential energies always decreases with time due to friction

and other dissipative effects. Since energy cannot dissipate for ever, the system eventually settles down to a minimum energy state. This is the behavior after which asymptotic stability was modeled. A good example is a simple mass-spring-damper system. The second method called *Lyapunov's indirect method* is based on linearizing the nonlinear system and studying the properties of the linearization. We will consider this method in more detail in the next chapter.

Consider the continuous-time system in (9.1). Let x_e be an equilibrium point and N be an open neighborhood of x_e.

Definition 9.3.1 (Lyapunov function) *A continuously differentiable function $V : N \to \mathbb{R}$ is a Lyapunov function for the system (9.1) if and only if it has the following two properties:*

1. *Positive Definiteness: $V(x) > 0$ for all $x \neq x_e$ in N*
2. *Dissipativeness: $\dot{V}(x) \leq 0$ in N. That is, the rate of change of V along the solutions of the system (9.1) emanating from points in N is less than zero for all time.* ■

Example 9.3.1 *Consider the system $\dot{x} = -x$. The function V defined as $V(x) = x^2$ is a Lyapunov function for this system. The system $\dot{x} = x$ has no Lyapunov function.* ■

Lyapunov functions are extremely difficult to construct for general nonlinear systems. We must search over all positive definite functions to find one that is dissipative. The set of functions that are positive definite is of infinite dimension making numerical search usually difficult. Even when physical notions of energy are known as in mechanical systems, a great deal of intuition is required to come up with the correct function. This is because mechanical energy is not always the Lyapunov function. But, a great simplification occurs in the case of LTI systems. In this case, the two methods of Lyapunov are identical and we can eliminate functions of higher degree from the search for Lyapunov function. We discuss this below.

Definition 9.3.2 (Quadratic Lyapunov function) *A Lyapunov function of the form*

$$V(x) = x^* P x$$

where P is a positive definite matrix is called a quadratic Lyapunov function. The matrix P is called the corresponding Lyapunov matrix. ■

Theorem 9.3.1 (Quadratic Lyapunov functions and LTI systems) *Consider the LTI system:*

$$\dot{x} = Ax$$

The following statements are equivalent.

1. *The LTI system is asymptotically stable.*

2. *The LTI system has a quadratic Lyapunov function.* ■

So, in the LTI case, we do not have to search over all positive definite functions. It is enough to consider quadratic Lyapunov functions. The next chapter will show how to construct such functions by solving linear equations.

Chapter 10

Lyapunov Stability of LTI Systems

Stability as defined in the last chapter is a system theoretic property. It deals with the behavior of system trajectories. An important feature of Lyapunov's direct and indirect approaches is that stability is determined without actually solving the system equations for the trajectories. Stability properties are deduced from other characteristics of the system. A complete study of stability is possible using these methods for the LTI system:

$$\dot{x} = Ax, \quad x(0) = x_0 \tag{10.1}$$

More importantly, we shall show that stability is equivalent to certain linear algebraic properties of the matrix A.

10.1 Lyapunov equation & inequality

Let A and Q be given $n \times n$ matrices. An equation of the form

$$AP + PA^* + Q = 0$$

is called the *continuous-time Lyapunov equation*. Here, P is the solution. The *discrete-time Lyapunov equation* has the form:

$$P - APA^* = Q$$

Lyapunov equations are linear in the (unknown) P. As with any linear system of equations, Lyapunov equations may or may not have solutions. For example, if $A = 0$ and $Q = 1$, there exists no solution for the continuous-time Lyapunov equation. On the other hand, if $A = 0$ and $Q = 0$, any scalar is a solution of the continuous-time Lyapunov equation. The following result shows when a solution exists.

Theorem 10.1.1 (Existence of solutions of Lyapunov equations) *Let A be a given $n \times n$ complex matrix. The following statements are true.*

1. *For each $n \times n$ complex matrix Q, the continuous-time Lyapunov equation*

$$AP + PA^* + Q = 0$$

has an unique solution P if and only if

$$\lambda_i + \overline{\lambda}_j \neq 0$$

for each pair of eigenvalues λ_i, λ_j of A.

2. *For each $n \times n$ complex matrix Q, the discrete-time Lyapunov equation*

$$P - APA^* = Q$$

has an unique solution P if and only if

$$\lambda_i \overline{\lambda}_j \neq 1$$

for each pair of eigenvalues λ_i, λ_j of A. ■

Let A be a real matrix. Then λ is an eigenvalue of A if and only if $\overline{\lambda}$ is an eigenvalue of A. Therefore, in this case, statement 1 of the theorem says that the Lyapunov equation has a unique solution if and only if the eigenvalues of A are distributed in the complex plane in such a way that no two of them add up to zero, i.e, they are not distributed symmetrically about the imaginary axis. Similar statements about the distribution of eigenvalues can be made in the case when A is complex and in the discrete-time case.

Suppose that the eigenvalues of A are contained in the open left half plane, i.e

$$\lambda_i(A) \in \{s : \text{ real part of } s < 0\}$$

Pick any two eigenvalues of A, say λ_1 and λ_2. They are in general complex numbers and can be written as:

$$\lambda_1 = \sigma_1 + j\omega_1 \quad \text{and} \quad \lambda_2 = \sigma_2 + j\omega_2$$

where the real parts σ_1 and σ_2 are both strictly negative. So,

$$\lambda_1 + \overline{\lambda}_2 = \sigma_1 + j\omega_1 + \sigma_2 - j\omega_2 = (\sigma_1 + \sigma_2) + j\,(\omega_1 - \omega_2)$$

cannot equal 0 because the real part is the sum of two numbers that are both strictly negative. Thus, any A whose eigenvalues are contained in the left half plane satisfies the requirement of statement 1 of Theorem 10.1.1 for the existence of a unique solution. We can actually say more as the following theorem shows.

Theorem 10.1.2 (Explicit formula for the solution of Lyapunov equation) *Let $A \in \mathbb{C}^{n\times n}$ and $Q \in \mathbb{C}^{n\times n}$ be given. The following statements are true.*

1. *Suppose that the eigenvalues of A are contained in the open left half plane. Then,*

$$P = \int_0^\infty e^{At} Q e^{A^* t} dt$$

exists as a matrix and is the unique solution of the continuous-time Lyapunov equation $AP + PA^* + Q = 0$.

2. *Suppose that the eigenvalues of A are contained in the open unit disk. Then,*

$$P = \sum_{k=0}^{\infty} A^k Q A^{*k}$$

exists as a matrix and is the unique solution of the discrete-time Lyapunov equation $P - APA^* = Q$.

■

These expressions for the solution are not meant for computations. We would compute a solution by solving the Lyapunov equation using linear algebra like any other equation. The formula is useful for theoretical applications and will be used frequently.

We now turn to Lyapunov inequality. Recall that a real matrix is positive if and only if it is symmetric and all its eigenvalues are positive. Similarly, a complex matrix is positive if and only if it is Hermitian and all its eigenvalues are positive. Now, given two positive matrices M and N, we say that M *is greater than or equal to* N (denoted by $M \geq N$) if and only if $M - N$ is positive. That is,

$$M \geq N \Leftrightarrow M - N \geq 0$$

We say that M *is less than or equal to* N (denoted by $M \leq N$) if $N - M$ is positive.

Let A be a given $n \times n$ matrix. An inequality of the form

$$AP + PA^* \leq 0$$

is called the *continuous-time Lyapunov inequality*. The *discrete-time Lyapunov inequality* has the form:

$$P - APA^* \leq 0$$

Solving these inequalities numerically are only a little more difficult than Lyapunov equations. But, they have advantages that far exceed those of equations and, as a result, are currently the preferred way to study stability and related properties.

10.2 Main stability theorems for continuous-time LTI systems

As the first of many algebraic characterizations of system-theoretic properties, we give testable necessary and sufficient conditions for stability in terms of the eigenvalues of the system matrix A. Some properties of stable systems will also be given. Recall that the LTI system (10.1) has the solution:

$$x(t) = e^{At} x_0$$

where x_0 is the initial condition.

Theorem 10.2.1 (Continuous-time case) *The following statements are true of LTI systems.*

1. *All solutions are bounded if and only if for each eigenvalue λ_i of A, we have:*
   a. *The real part of λ_i is ≤ 0 and*
   b. *If real part of λ_i is equal to zero, then its algebraic and geometric multiplicities are the same.*
2. *A system is stable about an equilibrium point if and only if all solutions are bounded.*
3. *If an equilibrium point is AS, then it is zero and it is the only equilibrium point.*
4. *A system is AS if and only if the real part of the eigenvalues of A are strictly less than zero (i.e. all the eigenvalues of A are in the open left half plane).*
5. *A system is AS if and only if e^{At} tends to zero as t tends to ∞.* ■

All the statements above can be proved by transforming to a new set of coordinates where the system matrix is the Jordan matrix associated with A. That is, introduce the similarity transformation $\widehat{x} = T^{-1}x$ where T is such that $\widehat{A} = T^{-1}AT$ and $\widehat{A}$ is Jordan. Note that the $\widehat{A}$ is: (a) upper-triangular, (b) the main diagonal contains the eigenvalues of A and (c) the diagonal above the main diagonal consists of 0s and 1s. As a result, e^{At} is: (a) upper-triangular, (b) the main diagonal contains $e^{\lambda(A)t}$, (c) the diagonal above contains 0s and $te^{\lambda(A)t}$, (d) the diagonal above the last one contains 0s and $t^2e^{\lambda(A)t}$ and so on.

Statements 1 and 2 of the theorem show when a system is stable. In order to check stability, we need to compute the eigenvalues of A as well as their geometric and algebraic multiplicities. This is an exceedingly difficult task from a numerical point of view. The third statement provides an easily testable condition for AS. In this case, we simply need to compute the eigenvalues of A and check if their real parts are strictly negative.

We now give other characterizations in terms of Lyapunov equations and inequalities. The difference between stability and AS apparent in the above theorem will make us consider the cases separately.

Theorem 10.2.2 (Continuous-time LTI asymptotic stability) *The following statements are equivalent.*

1. *The eigenvalues of A are contained in the open left half plane.*
2. *There exists $P > 0$ such that $AP + PA^* < 0$*
3. *There exists $Q > 0$ such that $QA + A^*Q < 0$*
4. *For each $R > 0$, there exists $P > 0$ such that $AP + PA^* + R = 0$*
5. *For each $S > 0$, there exists $Q > 0$ such that $QA + A^*Q + S = 0$* ■

Theorem 10.2.3 (Continuous-time LTI stability) *The following statements are equivalent.*

1. *LTI system is stable.*

2. *There exists* $P > 0$ *such that* $AP + PA^* \leq 0$. ■

10.3 Two stability related properties of systems

Suppose that a system is stable in some coordinate system. That is,

$$\dot{x} = Ax$$

and the eigenvalues of A are in the closed left half plane. Now for some reason, we dont like the coordinates x and wish to express the system in some other coordinate system z. It would be nice if stability notions were independent of coordinate systems. In other words, we would like stability to be a property of the system arising from the underlying physics rather than the coordinates chosen to express the system in. This turns out to be true.

Definition 10.3.1 (LTI coordinate transforms) *Consider an LTI system (10.1). Let* T *be an invertible matrix and define the matrix* $\widehat{A}$ *as:* $\widehat{A} = T^{-1}AT$. *The matrix* T *is called the coordinate (or similarity or Lyapunov) transformation,* $\widehat{A}$ *is said to be similar to* A, *the state vector*

$$\widehat{x} = T^{-1}x$$

is the transformed coordinates and the LTI system

$$\dot{\widehat{x}} = \widehat{A}\widehat{x}$$

is the transformed system. ■

The next theorem shows that stability is invariant under coordinate transformations.

Theorem 10.3.1 (Coordinate invariance) *Consider the systems*

$$\dot{x} = Ax \;\; \textit{and} \;\; \dot{\widehat{x}} = \widehat{A}\widehat{x}$$

where the coordinates x *and* $\widehat{x}$ *are related through a similarity transformation* T *(that is,* $\widehat{x} = T^{-1}x$*). Then,* $\dot{x} = Ax$ *is stable if and only if* $\dot{\widehat{x}} = \widehat{A}\widehat{x}$ *is stable. Further,* $\dot{x} = Ax$ *is asymptotically stable if and only if* $\dot{\widehat{x}} = \widehat{A}\widehat{x}$ *is asymptotically stable.* ■

Proof is easy and involves showing that eigenvalues and their multiplicities are invariant under similarity transforms.

Any reasonable model of a real system must have uncertainty which can be parametric or non-parametric. Consider the system:

$$\dot{x} = (A + \Delta A)\, x$$

where $A \in \mathbb{R}^{n\times n}$ is the *nominal* system model and ΔA is an unknown matrix representing parametric uncertainty (i.e. uncertainty in the elements of A). Usually, it is known that ΔA belongs to a collection of matrices $\mathbf{\Delta}$. For example, consider the second order system:

$$\dot{x} = \begin{bmatrix} -\zeta\omega_n & \omega_n \\ -\omega_n & -\zeta\omega_n \end{bmatrix} x$$

where the natural frequency ω_n is unknown, but belongs to the interval $[\omega_{min},\ \omega_{max}]$. Define

$$\mathbf{\Delta} = \left\{ \Delta A = \begin{bmatrix} -\zeta\omega & \omega \\ -\omega & -\zeta\omega \end{bmatrix} : \omega \in [\omega_{min},\ \omega_{max}] \right\}$$

and note that we can write the system as:

$$\dot{x} = (0 + \Delta A)\, x$$

where $\Delta A \in \mathbf{\Delta}$. We will say that the uncertain system is *robustly stable* if it is AS for every possible $\Delta A \in \mathbf{\Delta}$ (actually we need a little more). This means that no matter what the values of the actual parameters are, as long as they are in the admissible set $\mathbf{\Delta}$, the real system will be AS. The next result shows that every nominal system that is AS can admit some uncertainties, that is, it is robustly stable.

Theorem 10.3.2 (Robustness) *Suppose that the nominal system $\dot{x} = Ax$ is AS. Then, there exists $\epsilon > 0$ such that for any ΔA satisfying*

$$\|\Delta A\| < \epsilon$$

we have that the perturbed system

$$\dot{x} = (A + \Delta A)\, x$$

is AS. ■

10.4 Discrete-time LTI systems

As in the continuous-time case, stability of LTI discrete-time systems is intimately related to the location of eigenvalues of A. In the continuous-time case, eigenvalues must be in the (closed) left half plane for stability; whereas in the discrete-time case, eigenvalues must be in the (closed) unit disc for stability. The following theorem is the discrete-time analog of the main theorem of the previous section.

Theorem 10.4.1 *Consider the LTI system:*

$$x_{k+1} = Ax_k$$

The following statements are true.

1. *All solutions of the system are bounded if and only if for each eigenvalue λ_i of A, we have*
   a. *The magnitude of λ_i is ≤ 1 and*
   b. *If the magnitude of λ_i is equal to 1, then its algebraic and geometric multiplicities are the same.*
2. *The system is stable about an equilibrium point if and only if all solutions are bounded.*
3. *The system is AS if and only if the eigenvalues of A have magnitude strictly less than 1.*
4. *The system is AS if and only if A^k tends to zero as k tends to ∞.* ■

All of the results stated in the previous section have discrete-time analogs.

10.5 Lyapunov's indirect method

The following theorem explains why linear systems are important. It is known as *Lyapunov's first method* or *Lyapunov's indirect method*.

Theorem 10.5.1 *Consider the nonlinear system in (9.1) and let x_e be an equilibrium point. Assume that f is continuously differentiable in a neighborhood of x_e. Denote by*

$$\dot{x} = Ax$$

the linearization of (9.1) about x_e. The following statements are true:

1. *If the linearization is AS about 0, then the nonlinear system is LAS about x_e.*
2. *If the linearization is unstable about 0, then the nonlinear system is unstable about x_e.* ■

As an example of the usefulness of this theorem, consider the Van der Pol oscillator in Example 9.1.4. Linearization about the equilibrium point $(0, 0)$ gives:

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix} x$$

which can be shown to be unstable about $(0, 0)$. Hence, the Van der Pol oscillator is unstable about $(0, 0)$.

Note that the theorem allows us to infer stability of the nonlinear system *from* the stability of the linearization. The converse (stability of the linearization from the stability of nonlinear system) is not necessarily true as the following example illustrates.

Example 10.5.1 *The nonlinear system $\dot{x} = -x^3$ can be shown (show this) to be LAS about 0. But, the linearization about 0 is $\dot{x} = 0$ is not AS.* ■

Chapter 11

Controllability and Stabilizability

Consider the LTI system:

$$\dot{x} = Ax + Bu \tag{11.1}$$

where $x(t) \in \mathbb{R}^n$ is the state vector and $u(t) \in \mathbb{R}^m$ is the control input vector. In the previous chapters, u was simply referred to as the input vector; here in this chapter, it will be thought of as a control input. A control input is a vector of signals that the designer can choose in order to achieve some goal. This is unlike external disturbances and measurement noises which are not at our discretion (such signals are called *exogenous inputs*).

Recall that Lyapunov stability dealt with system behavior in the vicinity of an operating point. As circumstances change, we may have to transition from the current operating point to another, or more specifically, from one state to another. The problem of taking a system from one state to another in a safe and sound manner is a problem in control design. This is a hard problem and we will not attempt to solve it. Note that if there is no system trajectory passing through the initial and final states, then we cannot hope to transition at all. The controllability problem is to solve this easier problem of checking if there is a control input that takes the system from an initial state to a final state. It makes no claims about the safety and soundness of transition or about what happens upon reaching the final state.

11.1 Controllability

We begin with the definition of controllability. Recall that, given any initial condition $x(0) = x_0$, the system equation (11.1) can be solved to get:

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-\tau)}Bu(\tau)d\tau, \quad \forall\, t \geq 0 \tag{11.2}$$

where the first term is the initial condition (or zero input) response and the second term is the forced (or zero state) response.

Definition 11.1.1 (Controllable) *The LTI system (11.1) is controllable if and only if for any initial state $x_0 \in \mathbb{R}^n$ and final state $x_f \in \mathbb{R}^n$, there is a finite time $T > 0$ and a control input u defined in the interval $[0, T]$ such that*

$$x(T) = e^{AT}x_0 + \int_0^T e^{A(T-\tau)}Bu(\tau)d\tau \tag{11.3}$$

is equal to x_f. In other words, for any pair (x_0, x_f) of points, there is a control input that drives the system from the initial state x_0 to the final state x_f in finite time T. ■

A few things in the definition are very important. The time T taken to go from initial state to final state must be finite. We may take a second or a million years; but a finite amount of time. Also, a control input must exist for *any* pair of initial and final states (x_0, x_f) for the system to be controllable. The control input may not be unique, but at least one must exist. If a system is not controllable, then we say that it is *uncontrollable*.

Strictly speaking, in the above definition, we should say that the system is controllable during the time $[0, T]$. For LTI systems, it turns out that the system is controllable during some time interval $[0, \hat{T}]$ if and only if it is controllable for *any* (nonempty) time interval. This fact will become obvious from corollary 11.1.1. Therefore, the concept of controllability of LTI systems is independent of how long it takes, and we say that the system is controllable or uncontrollable. Also, from now on, we will say that *the pair (A, B) is controllable (uncontrollable)* to mean that the system $\dot{x} = Ax + Bu$ is controllable (uncontrollable). This terminology has a more algebraic flavor and is in tune with our theme of algebraic characterization of system properties.

Controllability problem requires us to find a control input that will take the system from an initial state to a final state in T seconds. Here, the time T, the initial state x_0 and the final state x_f are given. The unknown is the control input u which we must determine by solving the convolution integral equation in (11.3). The convolution integral is a linear operator in the control input and, as with all linear equations, we ask the following questions:

1. Given a final time T and a pair of initial and final states (x_0, x_f), does there exist a control input that drives the system from x_0 at $t = 0$ to x_f at $t = T$?

2. Does there exist a control input that drives the system from x_0 to x_f in finite time *for any pair of initial and final states* (x_0, x_f) ? That is, when is the system controllable ?

3. If a control exists, find all control inputs that drives the system from x_0 at $t = 0$ to x_f at $t = T$?

4. When is the control input unique ?

Compare these questions with the questions for the linear equations discussed prior to Theorem 3.4.2. As preparations, we introduce a number of definitions.

Definition 11.1.2 (Controllability matrix) *Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. The $n \times nm$ matrix*

$$Q_c = [\, B \quad AB \quad A^2B \quad \cdots \quad A^{n-1}B \,]$$

is called the controllability matrix of the pair (A, B). ■

Definition 11.1.3 (Controllable subspace) *Range of the controllability matrix Q_c is called the controllable subspace. That is, it is the set of all points $y \in \mathbb{R}^n$ for which the linear equation:*

$$Q_c v = y$$

has a solution $v \in \mathbb{R}^{nm}$. ■

Definition 11.1.4 (Finite-time controllability gramian) *Let $\infty > T > 0$. The matrix*

$$P_c(T) = \int_0^T e^{At} BB^* e^{A^* t} dt$$

is called the finite-time controllability gramian of the pair (A, B). ■

Since T is finite, this matrix always exists. In chapter 10, we saw

$$\int_0^\infty e^{At} M e^{A^* t} dt$$

which looks similar to the controllability gramian. In fact, the above matrix is called the *infinite-time controllability gramian* of the pair $(A, M^{1/2})$ where A is stable. Note also that with a change of variable, we can write $P_c(T)$ as

$$P_c(T) = \int_0^T e^{A(T-s)} BB^* e^{A^*(T-s)} ds$$

The next lemma is very important. It is similar to something we have seen before. In theorem 8.1.2, we considered the linear system $\dot{x} = Ax$. For this system, if the initial condition lies in an eigen-subspace of A, then the initial condition response will remain in that subspace for all time. Thus, eigen-subspaces are invariants for the system dynamics with no inputs. Here is the analog of an invariant subspace for the system dynamics with inputs.

Lemma 11.1.1 (Invariance of controllable subspace) *Suppose that $x_0 = 0$. Then, for any control input u, the solution*

$$x(t) = \int_0^T e^{A(t-\tau)} Bu(\tau) d\tau$$

satisfies $x(t) \in R(Q_c)$ for all $t \geq 0$. That is, for each $t \geq 0$, there exist $v_0(t), \cdots, v_{n-1}(t)$ such that:

$$x(t) = \sum_{k=0}^{n-1} A^k B v_i(t) = Q_c v(t)$$

where $v(t)$ is the vector obtained by stacking $v_0(t), \cdots, v_{n-1}(t)$ one below the other into a column. ■

Note that $x_0 = 0$ is in the range of Q_c because the range of a matrix is a subspace. So, if a solution starts at 0 which is a point in the controllable subspace, then it stays in the range of Q_c, equivalently, in the controllable subspace for all time. We chose $x_0 = 0$ for simplicity. In fact, if a solution starts at any point in the controllable subspace, then it stays in the controllable subspace for all time and for any control input. This implies that we cannot access any point outside the controllable subspace from within. We are now ready to state the answer to some of the questions raised earlier.

Theorem 11.1.1 (Solution of the controllability problem) *Let x_0, x_f and $T > 0$ be given. The following statements are equivalent:*

1. *There exists a control input u defined in the interval $[0, T]$ that drives the system from x_0 to x_f in time T.*
2. *$x_f - e^{AT}x_0$ is in the controllable subspace.*

Moreover, if statement 2 holds, then a control input u that takes x_0 to x_f is given by:

$$u(t) = B^* e^{A^*(T-t)} y, \quad \forall\, t \in [0, T]$$

where y is any solution of

$$P_c(T)y = x_f - e^{AT}x_0$$

and $P_c(T)$ is the finite-time controllability gramian. ■

Using this theorem, a control input that drives the system from x_0 to x_f in finite time can be computed in the following steps:

(1) Fix a final time $T > 0$ (any strictly positive number will do). Compute the controllability gramian

$$P_c(T) = \int_0^T e^{At} B B^* e^{A^* t} dt$$

and

$$z = x_f - e^{AT}x_0$$

(2) Check if the linear system:

$$P_c(T)y = z$$

has a solution. If so, find one. Use theorem 3.4.2 of chapter 3 for solving this linear algebra problem.

(3) Define the control input as in the previous theorem.

The linear system in Step 2 may or may not have a solution. If it has no solution, then there exists no control input that drives the system from x_0 to x_f. On the other hand, if it has many solutions, then each solution

will in general give rise to a different control input that drives the system from x_0 to x_f in T seconds along possibly different trajectories. If we compute

$$y = P_c(T)^+ z$$

where $P_c(T)^+$ is the Moore-Penrose inverse of $P_c(T)$, then the resulting control input is *the control input of least energy (in the $\mathcal{L}_2$ ()) sense)*. Finally, note that we can always define y as above and calculate a control input u whether or not the linear system in Step 2 has a solution. If there is no control input that drives the system from x_0 to x_f, then the control input calculated above brings the system from x_0 to a point as close as possible to x_f in T seconds.

We now ask when does there exist a control input for any pair of initial and final states, equivalently, when is the pair (A, B) controllable.

Corollary 11.1.1 (System controllability) *The pair (A, B) is controllable if and only if the range of Q_c is $\mathbb{R}^n$, equivalently, the rank of Q_c is n.* ■

Accordingly, the system can be taken from any initial state to any final state in finite time if and only if all the states are in the controllable subspace of the system. This makes sense intuitively. After all, by the invariance lemma 11.1.1, the controllable subspace cannot be left once inside. So, there can be no point outside it if the system is controllable. Notice that the testable condition for controllability given in the corollary is algebraic. We just need to determine the rank of the controllability matrix to check if the system is controllable or not. The next theorem gives several other algebraic characterizations of controllability.

Theorem 11.1.2 (Algebraic tests for controllability) *Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ be given matrices. The following statements are equivalent:*

1. *The pair (A, B) is controllable*
2. *Range of Q_c = $\mathbb{R}^n$ (Rank of Q_c = n)*
3. *Null space of Q_c^* = $\{0\}$*
4. *$P(T) > 0$ for all $T > 0$*
5. *Rank of*
$$[\,\lambda I - A \quad B\,]$$
is n for all complex numbers λ (this is known as PBH test).
6. *Rank of*
$$[\,\lambda I - A \quad B\,]$$
is n for all eigenvalues λ of A (this is known as modified PBH test).

7. *Suppose that there exists x such that $A^*x = \lambda x$ and $B^*x = 0$ for some complex number λ. Then, $x = 0$.*

8. *Suppose that (λ, x) is an eigenvalue-eigenvector pair of A^*. Then, $B^*x \neq 0$.*

9. *Given any set $\mu_1, \mu_2, \cdots, \mu_n$ of n complex numbers, there exists a state feedback gain K such that the eigenvalues of $A + BK$ are $\mu_1, \mu_2, \cdots, \mu_n$ (this is known as pole placement test).* ■

The last statement refers to the pole placement control problem which is to find a control law of the form:

$$u = Kx$$

called *state feedback control law* such that the eigenvalues of the closed loop system

$$\dot{x} = (A + BK)x$$

are at some pre-specified complex numbers. Usually, the poles of the closed loop system are chosen to lie in the open left half plane so as to realize good transient and steady state performances such as rise time, overshoot and zero tracking error. According to the statement, poles of the closed loop system can be placed *anywhere* if and only if the pair (A, B) is controllable.

We conclude this section with the following remarks. Controllability does not say anything about what happens to the system after T seconds. In fact, if the simulation is continued beyond T seconds, the system will continue to evolve and most likely leave the final state. This is one of the reasons why we mentioned in the introduction that controllability does not solve the control problem of transitioning from one operating condition to another. Also, we did not place any restrictions on the control input that drives the system from initial state to final state. Large and fast control inputs cannot be applied in practical systems due to rate and position saturation. So, it is unrealistic to go from initial to final states in arbitrarily small time.

11.2 Stabilizability

We now introduce a concept that is closer to control design than controllability is. Recall that the modified PBH test for controllability of the pair (A, B) says that the pair is controllable if and only if the matrix

$$[\,\lambda I - A \quad B\,] \tag{11.4}$$

has rank n for all eigenvalues λ of A. Suppose that the pair (A, B) is uncontrollable. Then, there exists an eigenvalue λ of A for which the rank of the matrix in (11.4) is strictly less than n. This allows us to classify eigenvalues of A into two groups.

Definition 11.2.1 (Controllable and uncontrollable modes) *An eigenvalue λ of A is called a controllable mode of the pair (A, B) if and only if the matrix defined in (11.4) has full rank. Otherwise, λ is called an uncontrollable mode of (A, B).* ■

Controllability implies that all the eigenvalues of A are controllable modes. This is too much to ask of real systems where usually only a fraction of the modes are controllable. Notice that, by the pole placement test of Theorem 11.1.2, if a system is controllable, then a state feedback law can be found to assign poles of the closed loop system to any desired location and thereby achieve any level of performance. Physical systems come with features that limit achievable performance. One of these features is the presence of uncontrollable modes.

Definition 11.2.2 (Stabilizable) *A pair (A, B) is (continuous-time) stabilizable if the real part of each uncontrollable mode of the pair (A, B) is strictly less than 0.* ■

This definition has a nice interpretation using pole placement. As mentioned before, if the pair (A, B) is controllable, then the poles of $A + BK$ can be placed anywhere. Thus, a controllable mode can be moved around in the complex plane as we wish using state feedback. If a mode is uncontrollable, then it cannot be moved by constant gain state feedback. Stabilizable means that all those modes that cannot be moved must be stable. This is the minimum requirement for the existence of a gain K so that the closed loop system $\dot{x} = (A + BK)x$ is asymptotically stable.

The main result on stabilizability is the following (compare with the corresponding statements for controllability):

Theorem 11.2.1 (Algebraic tests for stabilizability) *Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ be given. The following statements are equivalent.*

1. *The pair (A, B) is stabilizable*
2. *PBH Test: Rank of*
$$[\,\lambda I - A \quad B\,]$$
is n for all complex numbers λ in the closed right half plane.
3. *Modified PBH Test: Rank of*
$$[\,\lambda I - A \quad B\,]$$
is n for all eigenvalues λ of A in the closed right half plane.
4. *There exists K such that all the eigenvalues of $A + BK$ have strictly negative real parts.* ■

The test of stabilizability examines only the eigenvalues of A that are in the closed right half plane; whereas the corresponding test for controllability looks at all the eigenvalues of A. So, a controllable pair is stabilizable. The converse is not true.

Chapter 12

Observability and Detectability

Consider the LTI system:

$$\dot{x} = Ax, \qquad x(0) = x_0 \tag{12.1a}$$

$$y = Cx \tag{12.1b}$$

where x is the state vector and y is the output (measurement) vector. This system has no input and is driven by the initial state x_0. The state and output responses are:

$$x(t) = e^{At}x_0 \quad \text{and} \quad y(t) = Ce^{At}x_0$$

for all $t \geq 0$. So, if the initial state is known then the state $x(t)$ and the output $y(t)$ are known for all time $t \geq 0$. The subject of this chapter is the inverse problem of determining the initial condition x_0 from the output time history. If the initial condition can be uniquely determined, then

$$x(t) = e^{At}x_0$$

gives the evolution of states for all time. Thus, from the output time history, we obtain a complete picture of the evolution of internal system states.

Recall that the controllability problem involves finding control inputs to transition from one state to another in finite time. So, in some sense, it has to do with how effective the control channels are. Observability has to do with how effective the sensors are. Remember that a state is a representation of the internal workings of a system. Unfortunately, we cannot measure every state in practical applications either because there are too many states or because it is not physically possible. So, from a limited number of sensor read-outs which are algebraic combinations of states ($y = Cx$), we must determine exactly what is going on inside the system. This is the observability problem in a nutshell. It is a simpler problem than the practical problem of sensor selection and placement to observe a system.

The practical problem is also complicated by the presence of measurement noise which is neglected here. Noise will always prevent us from knowing the system states exactly. But, in applications, exact knowledge

of the states is not required. We only need to know an *estimate* of the state. This problem of estimating states from noisy measurements is known as the *state estimation* or *filtering* problem which has been extensively studied. Although we will not discuss filtering in this chapter, it should be noted that the concepts of observability and detectability are essential for filtering.

12.1 Observability

Definition 12.1.1 (Observable) *The LTI system (12.1) is observable if and only if there exists a finite time $T > 0$ such that $y(t) = 0$ for all $t \in [0, T]$ implies that $x_0 = 0$. That is, the system is observable if and only if the only initial condition that results in zero output over a nonempty time interval is the zero initial condition.*

The definition considers only a specific output response, namely, zero output. This may look restrictive; but by linearity, the definition is equivalent to the following: Suppose that y_1 and y_2 are outputs generated with initial states x_1 and x_2. Then, $y_1(t) = y_2(t)$ over a nonempty interval implies that $x_1 = x_2$. The main reason for stating the definition in terms of $y = 0$ is that zero output provides *no information* about the internal states and is thus the worst output. If the initial condition (and hence the internal state evolution) can be uniquely determined even for the worst output, then it can be determined when the output is not identically zero. If a system is not observable, then we say that it is *unobservable*.

Strictly speaking, in the above definition, we should say that the system is observable during the time $[0, T]$. For LTI systems, it turns out that the system is observable during some time nonempty interval $[0, \hat{T}]$ if and only if it is observable for any nonempty time interval. So, we shall simply say that the system is observable (or unobservable). Also, from now on, we will say that *the pair (C, A) is observable (or unobservable)* to mean that the system in (12.1) is observable (or unobservable).

Definition 12.1.2 (Observability matrix) *Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{m\times n}$. The $nm \times n$ matrix*

$$Q_o = \begin{bmatrix} C \\ CA \\ CA^2 \\ \cdot \\ \cdot \\ \cdot \\ CA^{n-1} \end{bmatrix}$$

is called the observability matrix of the pair (C, A).

Definition 12.1.3 (Unobservable subspace) *Null space of the observability matrix Q_o is called the unobservable subspace. That is, unobservable subspace is the set of all points $x \in \mathbb{R}^n$ such that $Q_o x = 0$.*

Definition 12.1.4 (Finite-time observability gramian) *Let $0 < T < \infty$. The matrix*

$$P_o(T) = \int_0^T e^{A^*t} C^* C e^{At} dt$$

is called the finite-time observability gramian of the pair (C, A).

The matrix

$$P_o = \int_0^\infty e^{A^*t} C^* C e^{At} dt$$

which exists if A is stable is called the *infinite-time observability gramian* of the pair (C, A). As before, with a change of variable, we can write $P_o(T)$ as

$$P_o(T) = \int_0^T e^{A^*(T-s)} C^* C e^{A(T-s)} ds$$

Recall from the last chapter on controllability that if a system starts out in the controllable subspace, then it cannot leave that subspace no matter what control input is applied. The analogous question here is what happens if we start in the *unobservable subspace*.

Lemma 12.1.1 (Invariance of unobservable subspace) *Suppose that x_0 is in the unobservable subspace. Then, $y(t) = 0$ for all time $t \geq 0$.*

The lemma means the following. If a point in the null space of Q_o (that is, a point in the unobservable subspace) is chosen as the initial state, then the output will be zero for all time. Another way of saying this is that if the output is not zero for some time t, then the initial state is not in the unobservable subspace. We should note two things here. First, (and this is very important) there is no such thing as an observable subspace (hence we say not in the unobservable subspace). The second thing is that if $y(t_1) \neq 0$ for some time t_1, then there is a time interval containing t_1 where $y(t) \neq 0$. This follows from the fact that y is infinitely differentiable.

Theorem 12.1.1 (Solution of the observability problem) *Let the output y be given over a nonempty interval $[0, T]$. The following statements are equivalent:*

1. *There exists an initial state x_0 such that $y(t) = Ce^{At}x_0$ for all $t \in [0, T]$.*
2. *$\int_0^T e^{A^*t} C^* y(t) dt$ is in the range of Q_o^*.*

Moreover, if statement 2 holds, then all initial states x_0 that produce the output y are given by:

$$x_0 = v + P_o(T)^+ \int_0^T e^{A^*t} C^* y(t) dt$$

where v is an arbitrary vector in the null space of Q_o and $P_o(T)^+$ is the Moore-Penrose inverse of the observability gramian $P_o(T)$.

We now ask when is the initial state unique, or equivalently, when is the pair, (C, A) observable.

Corollary 12.1.1 (System observability) *The pair (C, A) is observable if and only if the null space of Q_o is $\{0\}$, or equivalently, the rank of Q_o is n.*

According to the corollary, the initial system state and its subsequent evolution can be uniquely determined from the outputs if and only if the unobservable subspace contains nothing other than 0. This is intuitively clear because by the invariance Lemma 12.1.1 any initial state in the unobservable subspace gives zero output. Suppose that an initial state x_0 generates the output time history y. Then, since $y = y + 0$ and by linearity of the system, any initial condition of the form $x_0 + v$ where v is in the unobservable subspace (and, hence, its contribution to the output is zero) also generates y as the output. So, for us to be able to uniquely determine the initial state, all initial conditions of the form $x_0 + v$ must collapse to the same point x_0. This means that the only element v in the unobservable subspace is 0.

As in the case of controllability, the testable condition in the corollary is algebraic. We simply need to check if the rank of the observability matrix is equal to n to determine if the system is observable or not. The next theorem presents a few more algebraic characterizations. Compare the statements with the corresponding statements for controllability in Chapter 11.

Theorem 12.1.2 (Algebraic tests for observability) *Let A and C be given matrices with A being $n \times n$. The following statements are equivalent.*

1. *The pair (C, A) is observable*
2. *Null space of Q_o = $\{0\}$ (Rank of Q_o = n)*
3. *Range of Q_o^* = $\mathbb{R}^n$*
4. *$P_o(T) > 0$ for all $T > 0$*
5. *Rank of*
$$\begin{bmatrix} \lambda I - A \\ C \end{bmatrix}$$
is n for all complex numbers λ (this is the PBH test)
6. *Rank of*
$$\begin{bmatrix} \lambda I - A \\ C \end{bmatrix}$$
is n for all eigenvalues λ of A (this is the modified PBH test).
7. *Suppose that there exists x such that $Ax = \lambda x$ and $Cx = 0$. Then, $x = 0$.*
8. *Given any set $\mu_1, \mu_2, \cdots, \mu_n$ of n complex numbers, there exists an observer gain L such that the eigenvalues of $A + LC$ are $\mu_1, \mu_2, \cdots, \mu_n$ (this is the pole placement test).*

The last statement refers to observer design problem which is to find an observer system:

$$\dot{\hat{x}} = A\hat{x} - L\left(y - C\hat{x}\right)$$

such that the poles of the observer error system:

$$\dot{e} = (A + LC)e$$

are at some pre-specified complex numbers. Here, the observer error $e = x - \hat{x}$. Usually, the poles of the error system are chosen to lie in the open left half plane so as to realize good transient and steady state performances. According to the statement, poles of the error system can be placed *anywhere* if and only if the pair (C, A) is observable.

12.2 Detectability

Recall that the modified PBH test says that the pair (C, A) is observable if and only if the matrix:

$$\begin{bmatrix} \lambda I - A \\ C \end{bmatrix} \tag{12.2}$$

has rank n for all eigenvalues λ of A. Suppose that the pair (C, A) is unobservable. Then, there exists an eigenvalue λ of A for which the rank of the matrix in (12.2) is strictly less than n. This matrix involving (C, A) allows us to put the eigenvalues of A into two groups:

Definition 12.2.1 (Observable and unobservable modes) *An eigenvalue λ of A is an observable mode of the pair (C, A) if and only if the matrix in (12.2) has rank n. A mode of (C, A) that is not observable is called an unobservable mode of the pair (C, A).*

Observability means that all the eigenvalues of A are observable modes of the pair (C, A). As in the case of controllability, this is too much to ask of practical systems. Recall that, by the pole placement test of Theorem 12.1.2, if the system is observable, then an observer system can be designed with any level of performance. Physical systems come with unobservable modes that limit the level of observer performance that can be realized.

Definition 12.2.2 (Detectable) *A pair (C, A) is (continuous-time) detectable if the real part of each unobservable mode of the pair (C, A) is strictly less than 0.*

This definition has a nice interpretation in terms of observer design. As we know, if a mode is observable, then it can be moved to any place in the complex plane during observer design. If a mode is unobservable, then it cannot be moved by feedback of output. So, for zero steady state observer error, all unobservable

modes must be stable. If the unobservable modes are stable, then their contribution to the initial condition response decays exponentially. Hence, the observer error can be made to go to zero asymptotically.

The main result on detectability is the following (compare with the corresponding statements for observability as well as stabilizability in Chapter 11):

Theorem 12.2.1 *[Algebraic tests for detectability] Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{m\times n}$. The following statements are equivalent.*

1. *The pair (C, A) is detectable*

2. *Rank of*
$$\begin{bmatrix} \lambda I - A \\ C \end{bmatrix}$$
is n for all complex numbers λ in the closed right half plane (this is the PBH test).

3. *Rank of*
$$\begin{bmatrix} \lambda I - A \\ C \end{bmatrix}$$
is n for all eigenvalues λ of A in the closed right half plane (this is the modified PBH test).

4. *There exists L such that all the eigenvalues of $A - LC$ are in the open left half plane.*

12.3 Duality

Even a cursory look at the algebraic characterizations of controllability given Chapter 11 and those of observability in the previous sections reveals many similarities. This is in spite of their different system-theoretic origins. Recall that controllability deals with driving the system from an initial state to a finite state by choosing a control input, whereas observability deals with finding the initial state from an observed output time history. As we shall see below, the algebraic characterizations imply a deeper connection known as duality.

Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Consider a pair (A, B) and the associated controllability matrix

$$Q_c(A, B) = [\, B \quad AB \quad A^2B \quad \cdots \quad A^{n-1}B \,]$$

where, for the purpose of exposition, we use $Q_c(A, B)$ indicating that Q_c depends on A and B. Define

$$\hat{A} = A^T \quad \text{and} \quad \hat{C} = B^T$$

and consider the pair $(\hat{C}, \hat{A})$. The observability matrix associated with this pair is

$$Q_o(\hat{C}, \hat{A}) = \begin{bmatrix} \hat{C} \\ \hat{C}\hat{A} \\ . \\ . \\ . \\ \hat{C}\hat{A}^{n-1} \end{bmatrix} = \begin{bmatrix} B^T \\ B^T A^T \\ . \\ . \\ . \\ B^T A^{n-1T} \end{bmatrix} = Q_o(B^T, A^T)$$

where again we use $Q_o(\hat{C}, \hat{A})$ to indicate the observablity matrix depends on $\hat{C}$ and $\hat{A}$. Note that

$$Q_c(A, B) = Q_o(B^T, A^T)^T$$

i.e., *the controllability matrix of a pair* (A, B) *of real matrices is the transpose of the observability matrix of the pair* (B^T, A^T). It is well-known that the transpose of a finite-dimensional real matrix is its adjoint and is a mapping between dual spaces. This is why controllability and observability are said to be dual of each other.

Theorem 12.3.1 (Duality between controllability and observability) *Let* $A \in \mathbb{R}^{n\times n}$ *and* $B \in \mathbb{R}^{n\times m}$. *The pair* (A, B) *is controllable if and only if the pair* (B^T, A^T) *is observable.*

Hence, any test for controllability of a pair is a test for observability of the dual pair and vice versa. It is a simple matter to verify that stabilizability and detectability are also duals of each other.

Bibliography

[1] L. Ahlfors, *Complex analysis*, 3rd Edition, McGraw-Hill, 1979.

[2] K. Hoffman and R. Kunze, *Linear algebra*, Prentice-Hall, 1961.

[3] A. Michel and C. Herget, *Applied algebra and functional analysis*, Dover, 1981.

[4] A. Naylor and G. Sell, *Linear operator theory in engineering and science*, Rinehart and Winston, 1971.

[5] G. Strang, *Linear algebra and its applications*, Academic, 1980.

[6] G. Golub and C. Van Loan, *Matrix computations*, Johns Hopkins, 1983.

[7] W. Rugh, *Linear system theory*, Prentice-Hall, 1996.